Published on DATE 2020 (https://past.date-conference.com/date20)

Goto Session:
- 1.1 Opening Session: Plenary, Awards Ceremony & Keynote Addresses
- UB01 Session 1
  - 2.1 Executive Session: Memories for Emerging Applications
  - 2.2 Hardware-assisted Secure Systems
- 2.3 Fueling the future of computing: 3D, TFT, or disruptive memories?
- 2.4 Challenges in Analog Design Automation & Security
- 2.5 Pruning Techniques for Embedded Neural Networks
- 2.6 Improving reliability and fault tolerance of advanced memories
- 2.7 Optimizing emerging applications for power-efficient computing
- UB02 Session 2
- 3.0 LUNCHTIME KEYNOTE SESSION
- 3.1 Special Session: Architectures for Emerging Technologies
- 3.2 Accelerating Design Space Exploration
- 3.3 EU/ESA projects on Heterogeneous Computing
- 3.4 Accelerating Neural Networks and Vision Workloads
- 3.5 Parallel real-time systems
- 3.6 NoC in the age of neural network and approximate computing
- 3.7 Augmented and Assisted Living: A reality
- UB03 Session 3
- 4.1 Interactive Presentations
- 4.2 Timing in System-Level Modeling and Simulation
- 4.3 EU Projects on Nanoelectronics with CMOS and alternative technologies
- 4.4 Some run it hot, others do not
- 4.5 Adaptation and optimization for real-time systems
- 4.6 Artificial Intelligence and Secure Systems
- 4.7 Future computing fabrics: security and design integration
- UB04 Session 4
- 5.1 Special Day on "Embedded AI": Tutorial Overviews
- 5.2 Machine Learning Approaches to Analog Design
- 5.3 Special Session: Secure Composition of Hardware Systems
- 5.4 New Frontiers in Formal Verification for Hardware
- 5.5 Model-Based Analysis and Security
- 5.6 Logic synthesis towards fast, compact, and secure designs
- 5.7 Stochastic Computing
- 5.8 Special Session: HLS for AI HW
- UB05 Session 5
- 6.1 Special Day on "Embedded AI": Emerging Devices, Circuits and Systems
- 6.2 Secure and fast memory and storage
- 6.3 Special Session: Modern Logic Reasoning Methods for Functional ECO
- 6.4 Microarchitecture to the rescue of memory
- 6.5 Efficient Data Representations in Neural Networks
- 6.6 From DFT to Yield Optimization
- 6.7 Safety and efficiency for smart automotive and energy systems
- UB06 Session 6
- 7.0 LUNCHTIME KEYNOTE SESSION
- UB07 Session 7
- 7.1 Special Day on "Embedded AI": Industry AI chips
- 7.2 Reconfigurable Systems and Architectures
- 7.3 Special Session: Realizing Quantum Algorithms on Real Quantum Computing Devices
- 7.4 Simulation and verification: where real issues meet scientific innovation
- 7.5 Runtime support for multi/many cores
- 7.6 Attacks on Hardware Architectures
- 7.7 Self-Adaptive and Learning Systems
- UB08 Session 8
- 8.1 Special Day on "Embedded AI": Neuromorphic chips and systems
- 8.2 We are all hackers: design and detection of security attacks
- 8.3 Optimizing System-Level Design for Machine Learning
- 8.4 Architectural and Circuit Techniques toward Energy-efficient Computing
- 8.5 CNN Dataflow Optimizations
1.1 Opening Session: Plenary, Awards Ceremony & Keynote Addresses

**Date:** Tuesday 10 March 2020

**Time:** 08:30 - 10:30

**Location / Room:** Amphithéâtre Dauphine

**Chair:** Giorgio Di Natale, DATE 20 General Chair, FR

**Co-Chair:** Cristiana Bolchini, DATE 20 Programme Chair, IT

### Time  | Label | Presentation Title | Authors
---|---|---|---
08:15 | 1.1.1 | WELCOME ADDRESSES | Giorgio Di Natale
08:20 | 1.1.2 | PRESENTATION OF AWARDS | Philippe Magarshack, STMicroelectronics, FR

---

**Abstract**

Industrial IoT (IIoT) Systems are now becoming a reality. IIoT is distributed by nature, encompassing many complementary technologies. IIoT systems are composed of sensors, actuators, a means of communication and control units, and are moving into the factories, with the Industry 4.0 generation. In order to operate concurrently, all these IIoT components will require a wide range of technologies, in order to maintain such system-of-systems in a full operational, coherent and secure state. We identify and describe the four key enablers for the Industrial IoT:

1. more powerful and diverse embedded computing, available on ST’s latest STM32 microcontrollers and microprocessors,
2. augmented by AI applications at the edge (in the end devices), whose development is becoming enormously simplified by our specialized tools,
3. a wide set of connectivity technology, either with complete System-on-chip, or ready-to-use modules, and
4. a scalable security offer, thanks to either integrated features or dedicated security devices.

We conclude with some perspective on the usage of Digital Twins in the IIoT.
<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1.4</td>
<td>UB01 Session 1</td>
<td>PLENARY KEYNOTE: OPEN PARALLEL ULTRA-LOW POWER PLATFORMS FOR EXTREME EDGE AI</td>
<td>Luca Benini, ETH Zurich, CH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abstract</td>
<td>Edge Artificial Intelligence is the new megatrend, as privacy concerns and networks bandwidth/latency bottlenecks prevent cloud offloading of sensor analytics functions in many application domains, from autonomous driving to advanced prosthetic. The next wave of &quot;Extreme Edge AI&quot; pushes aggressively towards sensors and actuators, opening major research and business development opportunities. In this talk I will give an overview of recent efforts in developing an Extreme Edge AI platform based on open source parallel ultra-low power (PULP) Risc-V processors and accelerators. I will then look at what comes next in this brave new world of hardware renaissance.</td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td>End of session</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date: Tuesday 10 March 2020</th>
<th>Time: 10:30 - 12:30</th>
<th>Location / Room: Booth 11, Exhibition Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>Presentation Title</td>
<td>Authors</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------</td>
<td>---------</td>
</tr>
<tr>
<td>UB01.1</td>
<td>FUZZING EMBEDDED BINARIES LEVERAGING SYSTEMC-BASED VIRTUAL PROTOTYPES</td>
<td>Vladimir Herdt1, Daniel Grosse2 and Rolf Drechsler2</td>
</tr>
<tr>
<td></td>
<td>Authors:</td>
<td>1DFKI, DE; 2University of Bremen / DFKI GmbH, DE</td>
</tr>
<tr>
<td></td>
<td>Abstract</td>
<td>Verification of embedded software (SW) binaries is very important. Mainly, simulation-based methods are employed that execute (randomly) generated test-cases on virtual prototypes (VPs). However, to enable a comprehensive VP-based verification, sophisticated test-case generation techniques need to be integrated. Our demonstrator combines state-of-the-art fuzzing techniques with SystemC-based VPs to enable a fast and accurate verification of embedded SW binaries. The fuzzing process is guided by the coverage of the embedded SW as well as the SystemC-based peripherals of the VP. The effectiveness of our approach is demonstrated by our experiments, using RISC-V SW binaries as an example.</td>
</tr>
<tr>
<td></td>
<td>More information ...</td>
<td></td>
</tr>
<tr>
<td>UB01.2</td>
<td>SKELETON: AN OPEN SOURCE EDA TOOL FLOW FROM HIERARCHY SPECIFICATION TO HDL DEVELOPMENT</td>
<td>Ivan Rodriguez, Guillem Cabo, Javier Barrera, Jeremy Giesen, Alvaro Jover and Leonidas Kosmidis, BSC / UPC, ES</td>
</tr>
<tr>
<td></td>
<td>Authors:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Large hardware design projects have high overhead for project bootstrapping, requiring significant effort for translating hardware specifications to hardware design language (HDL) files and setting up their corresponding development and verification infrastructure. Skeletor (<a href="https://github.com/jaquerinte/Skeletor">https://github.com/jaquerinte/Skeletor</a>) is an open source EDA tool developed as a student project at UPC/BSC, which simplifies this process, by increasing developer's productivity and reducing typing errors, while at the same time lowers the bar for entry in hardware development. Skeletor uses a C/verilog-like language for the specification of the modules in a hardware project hierarchy and their connections, which is used to generate automatically the require skeleton of source files, their development and verification testbenches and simulation scripts. Integration with KiCad schematics and support for syntax highlighting in code editors simplifies further its use. This demo is linked with workshop W05.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>More information ...</td>
<td></td>
</tr>
<tr>
<td>UB01.3</td>
<td>LAGARTO: FIRST SILICON RISC-V ACADEMIC PROCESSOR DEVELOPED IN SPAIN</td>
<td>Ivan Rodriguez, Cristobal Ramirez Lazo1, Julian Pavan Rivera1, Vatistas Kostalabros1, Carlos Rojas Morales1, Miquel Moreto1, Jaume Abella1, Francisco1, Cazorla1, Adrian Cristal1, Roger Figueras1, Alberto Gonzalez1, Carles Hernandez1, Cesar Hernandez2, Neiel Leyva2, Joan Marimon1, Ricardo Martínez3, Jonnatan Mendoza1, Francesc Moll4, Marco Antonio Ramirez2, Carlos Rojas1, Antonio Rubio4, Abraham Ruiz1, Nehir Sonmez1, Lluis Teres3, Osman Unsal3, Mateo Valero1, Ivan Vargas1 and Luis Vila2</td>
</tr>
<tr>
<td></td>
<td>Authors:</td>
<td>1BSC / UPC, ES; 2ICIC-IPN, MX; 3IMB-CNM (CSIC), ES; 4UPC, ES; 5BSC, ES</td>
</tr>
<tr>
<td></td>
<td>Abstract</td>
<td>Open hardware is a possibility that has emerged in recent years and has the potential to be as disruptive as Linux was once, an open source software paradigm. If Linux managed to lessen the dependence of users in large companies providing software and software applications, it is envisioned that hardware based on ISAs can open source can do the same in their own field. In the Lagarto tapeout four research institutions were involved: Centro de Investigación en Computación of the Mexican IPN, Centro Nacional de Microelectrónica of the CSIC, Universitat Politècnica de Catalunya (UPC) and Barcelona Supercomputing Center (BSC). As a result, many bachelor, master and PhD students had the chance to achieve real-world experience with ASIC design and achieve a functional SoC. In the booth, you will find a live demo of the first ASIC and prototypes running on FPGA of the next versions of the SoC and core.</td>
</tr>
<tr>
<td></td>
<td>More information ...</td>
<td></td>
</tr>
<tr>
<td>UB01.4</td>
<td>PARALLEL ALGORITHM FOR CNN INFERENCE AND ITS AUTOMATIC SYNTHESIS</td>
<td>Takashi Matsumoto, Yukio Miyasaka, Xinpei Zhang and Masahiro Fujita, University of Tokyo, JP</td>
</tr>
<tr>
<td></td>
<td>Authors:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Abstract</td>
<td>Recently, convolutional neural network (CNN) has surpassed conventional methods in the field of image processing. This demonstration shows a new algorithm to calculate CNN inference using processing elements arranged and connected based on the topology of the convolution. They are connected in mesh and calculate CNN inference in a systolic way. The algorithm performs the convolution of all elements with the same output feature in parallel. We demonstrate a method to automatically synthesize an algorithm, which simultaneously performs the convolution and the communication of pixels for the computation of the next layer. We show with several sizes of input layers, kernels, and strides and confirmed that the correct algorithms were synthesized. The synthesis method is extended to the sparse kernel. The synthesized algorithm requires fewer cycles than the original algorithm. There were the more chances to reduce the number of cycles with the sparser kernel.</td>
</tr>
<tr>
<td></td>
<td>More information ...</td>
<td></td>
</tr>
</tbody>
</table>
UB01.5 FASTERThERMSIM: FAST AND ACCURATE THERMAL SIMULATIONS FROM CHIPLETS TO SYSTEM
Authors: Yu-Min Lee, Chi-Wen Pan, Li-Rui Ho and Hong-Wen Chiong, National Chiao Tung University, TW
Abstract
Recently, owing to the scaling down of technology and 2.5D/3D integration, power densities and temperatures of chips have been increasing significantly. Though commercial and computational fluid dynamics tools can provide accurate thermal maps, they may lead to inefficiency in thermal-aware design with huge runtime. Thus, we develop the chip/package/system-level thermal analyzer, called FasThermsim, which can assist you to improve your design under thermal constraints in pre/post-silicon stages. In FasThermsim, we consider three heat transfer modes, conduction, convection, and thermal radiation. We convert them to temperature-independent terms by linearization methods and build a compact thermal model (CTM). By applying numerical methods to the CTM, the steady-state and transient thermal profiles can be solved efficiently without loss of accuracy. Finally, an easy-to-use thermal analysis tool is implemented for your design, which is flexible and compatible, with the graphic user interface.
More information ...

UB01.6 INTACT: A 96-CORE PROCESSOR WITH 6 CHIPLETS 3D-STACKED ON AN ACTIVE INTERPOSER AND A 16-CORE PROTOTYPE RUNNING GRAPHICAL OPERATING SYSTEM
Authors: Eric Guthmuller1, Pascal Vivet1, César Fuguet2, Yvain Thonnart1, Gaël Pillonnet2 and Fabien Clermidy1
1Université Grenoble Alpes / CEA List, FR; 2Université Grenoble Alpes / CEA-Leti, FR
Abstract
We present a demonstrator for our 96-cores cache coherent 3D processor and a first prototype featuring 16 cores. The demonstrator consists in our 16-cores processor running commodity operating systems such as Linux and NetBSD on a PC-like motherboard with user-friendly devices such as a HDMI display, keyboard and mouse. A graphical desktop is displayed, and the user will interact with it through the keyboard and mouse. The demonstrator is able to run parallel applications to benchmark its performance in terms of scalability. The main innovation of our processor is its scalable cache coherent architecture based on distributed L2-caches and adaptive L3-caches. Additionally, the energy consumption is also measured and displayed by reading dynamically from the monitors of power-supply devices. Finally we will also show open packages of the 3D processor featuring 6 16-core chiplets (28 nm FDSOI) on an active interposer (65 nm) embedding Network-on-Chips, power management and IO controllers.
More information ...

UB01.7 EEC: ENERGY EFFICIENT COMPUTING VIA DYNAMIC VOLTAGE SCALING AND IN-NETWORK OPTICAL PROCESSING
Authors: Ryosuke Matsuo1, Jun Shiomi1, Yutaka Masuda2 and Tohru Ishihara3
1Hyojo University, JP; 2Nagoya University, JP
Abstract
This poster demonstration will show results of our two research projects. The first one is on a project of energy efficient computing. In this project we developed a power management algorithm which keeps the target processor always running at the most energy efficient operating point by appropriately tuning the supply voltage and threshold voltage under a specific performance constraint. This algorithm is applicable to wide variety of processor systems including high-end processors and low-end embedded processors. We will show the results obtained with actual RISC processors designed using a 65nm technology. The second one is on a project of in-network optical computing. We show optical functional units such as parallel multipliers and optical neural networks. Several key techniques for reducing the power consumption of optical circuits will be also presented. Finally, we will show the results of optical circuit simulation, which demonstrate the light speed operation of the circuits.
More information ...

UB01.8 CATANIS: CAD TOOL FOR AUTOMATIC NETWORK SYNTHESIS
Authors: Davide Quaglia, Enrico Fraccaroli, Filippo Nevi and Sohail Mushqo, Università di Verona, IT
Abstract
The proliferation of communication technologies for embedded systems opened the way for new applications, e.g., Smart Cities and Industry 4.0. In such applications hundreds or thousands of smart devices interact together through different types of channels and protocols. This increasing communication complexity forces computer-aided design methodologies to scale up from embedded systems in isolation to the global inter-connected system. Network Synthesis is the methodology to automatically allocate functionality onto network nodes and define the communication infrastructure among them. This booth will demonstrate the functionality of a graphic tool for automatic network synthesis developed by the Computer Science Department of University of Verona. It allows to graphically specify the communication requirements of a smart space (e.g., its map can be considered) in terms of sensing and computation tasks together with a library of node types and communication protocols to be used.
More information ...

UB01.9 PRE-IMPACT FALL DETECTION ARCHITECTURE BASED ON NEUROMUSCULAR CONNECTIVITY STATISTICS
Authors: Giovanni Mezzina, Sardar Meboob Hussain and Daniela De Venuto, Politecnico di Bari, IT
Abstract
In this demonstration, we propose an innovative multi-sensor architecture operating in the field of pre-impact fall detection (PIFD). The proposed architecture jointly analyzes cortical and muscular involvement when unexpected slippages occur during steady walking. The EEG and EMG are acquired through wearable and wireless devices. The control unit consists of an STM32L4 microcontroller and a Simulink modeling. The C implements the EMG computation, while the cortical analysis and

UB01.10 JOINTER: JOINING FLEXIBLE MONITORS WITH HETEROGENEOUS ARCHITECTURES
Authors: Giacomo Valente1, Tiziana Fanni2, Carlo Sau3, Claudio Rubattu2, Francesca Palumbo2 and Luigi Pomante1
1Università degli Studi dell’Aquila, IT; 2Università degli Studi di Sassari, IT; 3Università degli Studi di Cagliari, IT
Abstract
As embedded systems grow more complex and shift toward heterogeneous architectures, understanding workload performance characteristics becomes increasingly difficult. In this regard, run-time monitoring systems can support on obtaining the desired visibility to characterize a system. This demo presents a framework that allows to develop complex heterogeneous architectures composed of programmable processors and dedicated accelerators on FPGA, together with customizable monitoring systems, keeping under control the introduced overhead. The whole development flow (and related prototypical EDA tools), that starts with the accelerators creation using a dataflow model, in parallel with the monitoring system customization using a library of elements, showing also the final joining, will be shown. Moreover, a comparison among different monitoring systems functionalities on different architectures developed on Zynq7000 SoC will be illustrated.
More information ...

12:30 End of session
Memories play a prime role in virtually every modern computing systems. While memory technology has been able to follow the aggressive trend of scaling and keep up with the most stringent demands, there exists new applications for which traditional memories struggle to deliver viable solutions. In this context, and more than ever, novel memory technologies are required. Identifying a close match between a killer application and a supporting emerging memory technology will ensure unprecedented capabilities and open durable new horizons for computing systems. In this executive session, we will explore specific cases where novel memories (OxRAM and SOT MRAM in particular) are opening new applications unachievable with standard memories.

**2.2 Hardware-assisted Secure Systems**

**Date:** Tuesday 10 March 2020  
**Location / Room:** Chamrousse

**Chair:** Prabhat Mishra, University of Florida, US  
**Co-Chair:** Kavun Elif Bilge, University of Sheffield, GB

This session covers state-of-the-art hardware-assisted techniques for secure systems such as random number generators, PUFs, and logic locking & obfuscation. In addition, novel detection methods for hardware Trojans are presented.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:30</td>
<td>2.1.1</td>
<td>RESISTIVE RAM AND ITS DENSE 3D INTEGRATION FOR THEN3XT 1,000X</td>
<td>Author: Subhashish Mitra, Stanford University, US</td>
</tr>
<tr>
<td>12:00</td>
<td>2.1.2</td>
<td>EMERGING MEMORIES FOR VON NEUMANN AND FOR NEUROMORPHIC COMPUTING</td>
<td>Author: Jamil Kawa, Synopsys, US</td>
</tr>
<tr>
<td>12:30</td>
<td>2.1.3</td>
<td>RERAM TECHNOLOGY FOR NEXT GENERATION AI AND COST-EFFECTIVE EMBEDDED MEMORY</td>
<td>Author: Amir Regev, Weebit Nano, AU</td>
</tr>
<tr>
<td>13:00</td>
<td></td>
<td>End of session</td>
<td></td>
</tr>
</tbody>
</table>

**2.2.1 BACKTRACKING SEARCH FOR OPTIMAL PARAMETERS OF A PLL-BASED TRUE RANDOM NUMBER GENERATOR**

**Speaker:** Brice Colombier, Université de Lyon, FR  
**Authors:** Brice Colombier¹, Nathalie Bochard¹, Florent BERNARD² and Lilian Bossuet ¹  
¹Université de Lyon, FR; ²Laboratory Hubert Curien, University of Lyon, UJM Saint-Etienne, FR

**Abstract**

The phase-locked loop-based true random number generator (PLL-TRNG) extracts randomness from clock jitter. It is an interesting construct because it comes with a stochastic model, making it certifiable by certification bodies. However, bringing it to good performance is difficult since it comes with multiple parameters to tune. This article proposes to use backtracking to determine these parameters. Compared to existing methods, based on genetic algorithms or exhaustive search of a feasible set of parameters, backtracking has several advantages. Indeed, since this method is expressible by constraint programming, it provides very good readability. Constraints can be specified in a very straightforward and maintainable way. It also exhibits good performance and generates PLL-TRNG configurations rapidly. Finally, it allows to integrate new exploratory design constraints for the PLL-TRNG very easily. We provide experimental results with a PLL-TRNG implemented on three FPGA families that come with different physical constraints, showing that the method allows to find good parameters for every one of them. Moreover, we were able to obtain configurations that lead to an increase 59% in throughput and 82% in jitter sensitivity on average, thereby generating random numbers of higher quality at a faster rate. This approach also paves the way for new design exploration strategies for PLL-TRNG. The source code of our implementation is open source and available online for reproducibility and reuse.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**2.2.2 LONG-TERM CONTINUOUS ASSESSMENT OF SRAM PUF AND SOURCE OF RANDOM NUMBERS**

**Speaker:** Rui Wang, Intrinsic-ID, NL  
**Authors:** Rui Wang, Georgios Selimis, Roel Maes and Sven Goossens, Intrinsic-ID, NL

**Abstract**

The qualities of Physical Unclonable Functions (PUFs) suffer from several noticeable degradations due to silicon aging. In this paper, we investigate the long-term effects of silicon aging on PUFs derived from the start-up behavior of Static Random Access Memories (SRAM). Previous research on SRAM aging is based on transistor-level simulation or accelerated aging test at high temperature and voltage to observe aging effects within a short period of time. In contrast, we have run a long-term continuous power-up test on 16 Arduino Leonardo boards under nominal conditions for two years. In total, we collected around 175 million measurements for reliability, uniqueness and randomness evaluations. Analysis shows that the number of bits that flip with respect to the reference increased by 19.3% while min-entropy of SRAM PUF noise improves by 19.3% on average after two years of aging. The impact of aging on reliability is smaller under nominal conditions than was previously assessed by the accelerated aging test. The test we conduct in this work more closely resembles the conditions of a device in the field, and therefore we more accurately evaluate how silicon aging affects SRAM PUFs.

**Download Paper (PDF; Only available from the DATE venue WiFi)**
In the post-CMOS era, the future of computing relies more and more on emerging technologies, like resistive memories, TFT and 3D integration or their combination, to continue performance improvements: from a novel accelerating solution for deep neural networks with ferroelectric transistor technology, to a physical design methodology for face-to-face 3D ICs to enable commercial-quality IC layouts.

At DATE, the 32nd IEEE/ACM International Conference on Computer-Aided Design, you will have the opportunity to listen to leading experts present their work on these and other topics.

**2.3 Fueling the future of computing: 3D, TFT, or disruptive memories?**

**Presentation Title:** RESCUING LOGIC ENCRYPTION IN POST-SAT ERA BY LOCKING & OBfuscation

**Authors:** Hai Zhou, Northwestern University, US

**Abstract:**
Recently, a defense is

**Talk Title:** SELECTIVE CONCOLIC TESTING FOR HARDWARE TROJAN DETECTION IN BEHAVIORAL SYSTEMC DESIGNS

**Authors:** Bin Lin, Portland State University, US

**Abstract:**
With the growing complexities of modern SoC designs and increasingly shortened time-to-market requirements, new design paradigms such as outsourced design services have emerged. Design abstraction level has also been raised from RTL to ESL. Modern SoC designs in ESL often integrate a variety of third-party behavioral intellectual properties, as well as utilizing EDA tools intensively, to improve design productivity. However, this new design trend makes modern SoCs more vulnerable to hardware Trojan attacks. Although hardware Trojan detection has been studied for more than a decade in RTL and lower levels, it has just gained attention recently in ESL designs. In this paper, we present a novel approach for generating test cases by selective concolic testing to detect hardware Trojans in ESL. We have evaluated our approach on an open source benchmark that includes various types of hardware Trojans. The experimental results demonstrate that our approach is able to detect hardware Trojans effectively and efficiently.

**Presentation Title:** TEST PATTERN SUPERPOSITION TO DETECT HARDWARE TROJANS

**Authors:** Alex Oralioiu, University of California, San Diego, US

**Abstract:**
Current methods for the detection of hardware Trojans inserted by an untrusted foundry are either accompanied by unreasonable costs in design/test pattern overhead, or return results that fail to provide confident trustability. The challenges faced by these side-channel techniques are primarily a result of process variation, which renders pre-silicon expectations nearly meaningless in predicting the behavior of a manufactured IC. To overcome this hindrance in a cost-effective manner, we propose an easy-to-implement test pattern-based approach that is self-referential in nature, capable of dissecting and understanding the characteristics of a given manufactured IC to hone in on aberrant measurements that are demonstrative of malicious Trojan hardware. By leveraging the superposition principle to cancel out non-Trojan noise, we can isolate and magnify Trojan circuit effects, all within a regime considerate of leveraging the superposition principle to cancel out non-Trojan noise, we can isolate and magnify Trojan circuit effects, all within a regime.

**Presentation Title:** DYNUMLOCK: UNLOCKING SCAN CHAINS OBfuscated USING DYNAMIC KEYS

**Authors:** Nimisha Limaye, New York University, US

**Abstract:**
Outsourcing in semiconductor industry opened up venues for faster and cost-effective chip manufacturing. However, this also introduced untrusted entities with malicious intent, to steal intellectual property (IP), overproduce the circuits, insert hardware Trojans, or counterfeit the chips. Recently, a defense is proposed to obfuscate the scan access based on a dynamic key that is initially generated from a secret key but changes in every clock cycle. This defense can be considered as the most rigorous defense among all the scan locking techniques. In this paper, we propose an attack that remolds this defense into one that can be broken by the SAT attack, while we also note that our attack can be adjusted to break other less rigorous (key that is updated less frequently) scan locking techniques as well.

End of session
Switching lattices consisting of four-terminal switches are introduced as area-efficient structures to realize logic functions. Many optimization algorithms have been proposed, including exact ones, realizing logic functions on lattices with the fewest number of four-terminal switches, as well as heuristic ones. Hence, Switching lattices consisting of four-terminal switches are introduced as area-efficient structures to realize logic functions. Many optimization algorithms have been proposed, including exact ones, realizing logic functions on lattices with the fewest number of four-terminal switches, as well as heuristic ones. Hence, Switching lattices consisting of four-terminal switches are introduced as area-efficient structures to realize logic functions. Many optimization algorithms have been proposed, including exact ones, realizing logic functions on lattices with the fewest number of four-terminal switches, as well as heuristic ones. Hence, Switching lattices consisting of four-terminal switches are introduced as area-efficient structures to realize logic functions. Many optimization algorithms have been proposed, including exact ones, realizing logic functions on lattices with the fewest number of four-terminal switches, as well as heuristic ones. Hence, Switching lattices consisting of four-terminal switches are introduced as area-efficient structures to realize logic functions. Many optimization algorithms have been proposed, including exact ones, realizing logic functions on lattices with the fewest number of four-terminal switches, as well as heuristic ones. Hence, Switching...
AN EFFICIENT BAYESIAN OPTIMIZATION APPROACH FOR ANALOG CIRCUIT SYNTHESIS VIA SPARSE GAUSSIAN PROCESS MODELING

Speaker: Biao He, Fudan University, CN
Authors: Biao He1, Shuhan Zhang1, Fan Yang2, Changhao Yan1, Dian Zhou3 and Xuan Zeng1

Abstract: Bayesian optimization with Gaussian process models has been proposed for analog synthesis, since it is efficient for the optimizations of expensive black-box functions. However, the computational cost for training and prediction of Gaussian process models are $\mathcal{O}(N^3)$ and $\mathcal{O}(N^2)$, respectively, where $N$ is the number of data points. The overhead of the Gaussian process modeling would be negligible as $N$ grows linearly with the problem size. Recently, a Bayesian optimization approach using neural network has been proposed to address this problem. It reduces the computational cost of training/prediction of Gaussian process models to $\mathcal{O}(N)$ and $\mathcal{O}(1)$, respectively. However, the proposed approach is still time-consuming for large problem sizes. In this paper, we propose a novel Bayesian optimization approach using a neural network problem (SPGP). The idea is to use $\mathcal{O}(N)$ instead of $\mathcal{O}(N^3)$ and $\mathcal{O}(N^2)$, respectively. However, it reduces the computational cost of training/prediction to $\mathcal{O}(N)$ and $\mathcal{O}(1)$, which is more efficient than the traditional Gaussian process. Several experiments were conducted to demonstrate the efficiency of the proposed approach.

Download Paper (PDF; Only available from the DATE venue WiFi)
exploit pruning in embedded processing architectures. The solutions presented extend the sparsity concept to the bit level with an enhanced bit-level pruning technique based on 

Network pruning has been applied successfully to reduce the computational and memory footprint of neural network processing. This session presents three innovations to better

Dirk Ziegenbein, Robert Bosch GmbH, DE
Marian Verhelst, KU Leuven, BE

Chair:
Location / Room:
Date:
2.5 Pruning Techniques for Embedded Neural Networks

Tuesday 10 March 2020
11:30 - 13:00

Byungmin Ahn and Taewhan Kim, Seoul National University, KR

Authors

Speaker

Presentation Title
DEEPER WEIGHT PRUNING WITHOUT ACCURACY LOSS IN DEEP NEURAL NETWORKS

Byungmin Ahn, Seoul National University, KR

Authors

Speaker

Presentation Title
DEEPER WEIGHT PRUNING WITHOUT ACCURACY LOSS IN DEEP NEURAL NETWORKS

Byungmin Ahn, Seoul National University, KR

Authors

Speaker

Presentation Title
DEEPER WEIGHT PRUNING WITHOUT ACCURACY LOSS IN DEEP NEURAL NETWORKS

Byungmin Ahn, Seoul National University, KR

Authors

Download Paper (PDF; Only available from the DATE venue WiFi)
2.6 Improving reliability and fault tolerance of advanced memories

Date: Tuesday 10 March 2020
Time: 11:30 - 13:00
Location / Room: Lesdiguières

Chair: Mourin Benabedbeni, TIMA Laboratory, FR
Co-Chair: Said Hamdioui, TU Delft, NL

This session discusses reliability issues for different memory technologies; addressing fault tolerance of memristors, how to reduce simulations with importance sampling and advance metrics as measure for the reliability of NAND flash memories.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00</td>
<td>2.5.2</td>
<td>FLEXIBLE GROUP-LEVEL PRUNING OF DEEP NEURAL NETWORKS FOR ON-DEVICE MACHINE LEARNING</td>
<td>Dongkun Shin, Sungkyunkwan University, KR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker: Dongkun Shin, Sungkyunkwan University, KR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Authors:</strong> Kwangbgee Lee, Hoseung Kim, Hayun Lee and Dongkun Shin, Sungkyunkwan University, KR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract:</strong> Network pruning is a promising compression technique to reduce computation and memory access cost of deep neural networks. Pruning techniques are classified into two types: fine-grained pruning and coarse-grained pruning. Fine-grained pruning eliminates individual connections if they are insignificant and thus usually generates irregular networks. Therefore, it can fail to reduce inference time. Coarse-grained pruning such as filter-level and channel-level techniques can make hardware-friendly networks. However, it can suffer from low accuracy. In this paper, we focus on the group-level pruning method to accelerate deep neural networks on mobile GPUs, where several adjacent weights are pruned in a group to mitigate the irregularity of pruned networks while providing high accuracy. Although several group-level pruning techniques have been proposed, the previous techniques select weight groups to be pruned at group-size-aligned locations to reduce the problem space. In this paper, we propose an unaligned approach to improve the accuracy of the compressed model. We can find the optimal solution of the unaligned group selection problem with dynamic programming. Our technique also generates balanced sparse networks to get load balance at parallel computing units. Experiments demonstrate that the 2D unaligned group-level pruning shows 3.12% lower error rate at ResNet-20 network on CIFAR-10 that compared to the previous 2D aligned group-level pruning under the 95% sparsity.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td>2.5.3</td>
<td>SPARSITY-AWARE CACHES TO ACCELERATE DEEP NEURAL NETWORKS</td>
<td>Vinod Ganesan, IIT Madras, IN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker: Vinod Ganesan, IIT Madras, IN; Pratyush Kumar, Purdue University, US</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Authors:</strong> Vinod Ganesan1, Sanchari Sen1, Pratyush Kumar1, Neel Gala1, Kamakoti Veezhinatha1 and Anand Raghunathan2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>1IIT Madras, IN; 2Purdue University, US</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract:</strong> Deep Neural Networks (DNNs) have transformed the field of artificial intelligence and represent the state-of-the-art in many machine learning tasks. There is considerable interest in using DNNs to realize edge intelligence in highly resource-constrained devices such as wearables and IoT sensors. Unfortunately, the high computational requirements of DNNs pose a serious challenge to their deployment in these systems. Moreover, due to tight cost (and hence, area) constraints, these devices are often unable to accommodate hardware accelerators, requiring DNNs to execute on the General Purpose Processor (GPP) cores that they contain. We address this challenge through lightweight micro-architectural extensions to the memory hierarchy of GPPs that exploit a key attribute of DNNs, viz. sparsity, or the prevalence of zero values. We propose SparseCache, an enhanced cache architecture that utilizes a null cache based on a Ternary Content Addressable Memory (TCAM) to compactly store zero-valued cache lines while storing non-zero lines in a conventional data cache. By storing address rather than values for zero-valued cache lines, SparseCache increases the effective cache capacity, thereby reducing the overall miss rate and execution time. SparseCache utilizes a Zero Detector and Approximator (ZDA) and Address Merger (AM) to perform reads and writes to the null cache. We evaluate SparseCache on four state-of-the-art DNNs programmed with the Caffe framework. SparseCache achieves 5-28% reduction in miss-rate, which translates to 5-21% reduction in execution time, with only 0.1% area and 3.8% power overhead in comparison to a low-end Intel Atom Z-series processor.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
<td></td>
</tr>
<tr>
<td>13:00</td>
<td>2.5.4</td>
<td>TFAPPROX: TOWARDS A FAST EMULATION OF DNN APPROXIMATE HARDWARE ACCELERATORS ON GPU</td>
<td>Filip Vaverka, Vojtech Mrazek, Zdenek Vasicek and Lukas Sekanina, Brno University of Technology, CZ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker: Zdenek Vasicek, Brno University of Technology, CZ</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Authors:</strong> Filip Vaverka, Vojtech Mrazek, Zdenek Vasicek and Lukas Sekanina, Brno University of Technology, CZ</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract:</strong> Energy efficiency of hardware accelerators of deep neural networks (DNN) can be improved by introducing approximate arithmetic circuits. In order to quantify the error introduced by using these circuits and avoid the expensive hardware prototyping, a software emulator of the DNN accelerator is usually executed on CPU or GPU. However, this emulation is typically two or three orders of magnitude slower than a software DNN implementation running on CPU or GPU and operating with standard floating point arithmetic instructions and common DNN libraries. The reason is that there is no hardware support for approximate arithmetic operations on common CPUs and GPUs and these operations have to be expensively emulated. In order to address this issue, we propose an efficient emulation method for approximate circuits utilized in a given DNN accelerator which is emulated on GPU. All relevant approximate circuits are implemented as look-up tables and accessed through a texture memory mechanism of CUDA capable GPUs. We exploit the fact that the texture memory is optimized for irregular read-only access and in some GPU architectures is even implemented as a dedicated cache. This technique allowed us to reduce the inference time of the emulated DNN accelerator approximately 200 times with respect to an optimized CPU version on complex DNNs such as ResNet. The proposed approach extends the TensorFlow library and is available online at <a href="https://github.com/ehw-fit/tf-approximate">https://github.com/ehw-fit/tf-approximate</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
<td></td>
</tr>
</tbody>
</table>
Memristor based crossbar (MBC) can execute neural network computations in an extremely energy efficient manner. However, stuck-at faults make memristors represent network weight incorrectly, thus degrading classification accuracy of the network deployed on the MBC significantly. By carefully analyzing all the possible fault combinations in a pair of differential crossbars, we found that most of the stuck-at faults can be accommodated perfectly by mapping a zero value weight onto the memristors. Based on such observation, in this paper we propose a target sparsifying based fault tolerant scheme for the MBC which executes neural network applications. We first exploit a heuristic algorithm to map weight matrix onto the MBC, aiming at minimizing weight variations in the presence of stuck-at faults. After that, some weights mapped onto the faulty memristors which still have large variations will be purposefully forced to zero value. Network retraining is then performed to recover classification accuracy. For a 4-layer CNN designed for MNIST digit recognition, experimental results demonstrate that our scheme can achieve almost no accuracy loss when 10% of memristors in the MBC are faulty. As the faulty memristors increasing to 20%, accuracy loss is only within 3%.

Download Paper (PDF; Only available from the DATE venue WiFi)
BeldPC: Bit Errors Aware Adaptive Rate LDPC Codes for 3D TLC NAND Flash Memory

Speaker:
Meng Zhang, Huazhong University of Science & Technology, CN

Authors:
Meng Zhang, Fei Wu, Qin Yu, Weihua Liu, Lanlan Cui, Yahui Zhao and Changsheng Xie, Huazhong University of Science & Technology, CN

Abstract
Three-dimensional (3D) NAND flash memory has high capacity and cell storage density by using the multi-bit technology and vertical stack architecture, but degrading data reliability due to high raw bit error rates (RBER) caused by program/erase (P/E) cycles and retention periods. Low-density parity-check (LDPC) codes become more popular error-correcting technologies to improve data reliability due to strong error correction capability, but introducing more decoding iterations at higher RBER. To reduce decoding iterations, this paper proposes BelDPC: bit errors aware adaptive rate LDPC codes for 3D triple-level cell (TLC) NAND flash memory. Firstly, bit error characteristics in 3D charge trap TLC NAND flash memory are studied on a real FPGA testing platform, including asymmetric bit flipping and temporal locality of bit errors. Then, based on these characteristics, a high-efficiency LDPC code is designed. Experimental results show BelDPC can reduce decoding iterations under different P/E cycles and retention periods.

Download Paper (PDF; Only available from the DATE venue WiFi)
FPGA-DSP: A PROTOTYPE FOR HIGH QUALITY DIGITAL AUDIO SIGNAL PROCESSING BASED ON AN FPGA

Authors:
Bernhard Riess and Christian Epe, University of Applied Sciences Düsseldorf, DE

Abstract
Our demonstrator presents a prototype of a new digital audio signal processing system which is based on an FPGA. It achieves a performance that up to now has been preserved to costly high-end solutions. Main components of the system are an analog/digital converter, an FPGA to perform the digital signal processing tasks, and a digital/analog converter implemented on a printed circuit board. To demonstrate the quality of the audio signal processing, infinite impulse response, finite impulse response filters and a delay effect were realized in VHDL. More advanced signal processing systems can easily be implemented due to the flexibility of the FPGA. Measured results were compared to state of the art audio signal processing systems with respect to size, performance and cost. Our prototype outperforms systems of the same price in quality, and outperforms systems of the same quality at a maximum of 20% of the price. Examples of the performance of our system can be heard in the demo.

More information ...
UB02.5 AT-SPEED DFT ARCHITECTURE FOR BUNDLED-DATA CIRCUITS
Authors: Riccardo Aquino Guazzelli and Laurent Fesquet, Université Grenoble Alpes, FR
Abstract
At-speed testing for asynchronous circuits is still an open concern in the literature. Due to its timing constraints between control and data paths, Design for Testability (DFT) methodologies must test both control and data paths at the same time in order to guarantee the circuit correctness. As Process Voltage Temperature (PVT) variations significantly impact circuit design in newer CMOS technologies and low-power techniques such as voltage scaling, the timing constraints between control and data paths must be tested after fabrication not only under nominal conditions but through a range of operating conditions. This work explores an at-speed testing approach for bundled data circuits, targeting the micropipeline template. The main target of this test approach focuses on whether the sized delay lines in control paths respect the local timing assumptions of the data paths.
More information ...

UB02.6 INTACT: A 96-CORE PROCESSOR WITH 6 CHIPLETS 3D-STACKED ON AN ACTIVE INTERPOSER AND A 16-CORE PROTOTYPE RUNNING GRAPHICAL OPERATING SYSTEM
Authors: Eric Guthmuller1, Pascal Vivet1, César Fuguet1, Yvain Thonnart2, Gaël Pillonnet3 and Fabien Clermidy1
1Université Grenoble Alpes / CEA List, FR; 2Université Grenoble Alpes / CEA-Leti, FR
Abstract
We built a demonstrator for our 96-cores cache coherent 3D processor and a first prototype featuring 16 cores. The demonstrator consists in our 16-cores processor running commodity operating systems such as Linux and NetBSD on a PC-like motherboard with user-friendly devices such as a HDMI display, keyboard and mouse. A graphical desktop is displayed, and the user will interact with it through the keyboard and mouse. The demonstrator is able to run parallel applications to benchmark its performance in terms of scalability. The main innovation of our processor is its scalable cache coherent architecture based on distributed L2-caches and adaptive L3-caches. Additionally, the energy consumption is also measured and displayed by reading dynamically from the monitors of power-supply devices. Finally we will also show open packages of the 3D processor featuring 6 16-core chiplets (28 nm FDSOI) on an active interposer (65 nm) embedding Network-on-Chips, power management and I/O controllers.
More information ...

UB02.7 GENERATING ASYNCHRONOUS CIRCUITS FROM CATAPULT
Authors: Yoan Decoduri1, Jean Simatic2, Katell Morin-Allory3 and Laurent Fesquet3
1University Grenoble Alpes, FR; 2HaWaIL.Tech, FR; 3Université Grenoble Alpes, FR
Abstract
In order to spread asynchronous circuit design to a large community of designers, High-Level Synthesis (HLS) is probably a good choice because it requires limited design technical skills. HLS usually provides an RTL description, which includes a data-path and a control-path. The desynchronization process is only applied to the control-path, which is a Finite State Machine (FSM). This method is sufficient to make asynchronous the circuit. Indeed, data are processed step by step in the pipeline stages, thanks to the desynchronized FSM. Thus, the data-path computation time is no longer related to the clock period but rather to the average time for processing data into the pipeline. This tends to improve speed when the pipeline stages are not well-balanced. Moreover, our approach helps to quickly design data-driven circuits while maintaining a reasonable cost, a similar area and a short time-to-market.
More information ...

UB02.8 WALLANCE: AN ALTERNATIVE TO BLOCKCHAIN FOR IOT
Authors: Laic Dalmasso, Florent Bruguet, Pascal Benoit and Achraf Lamlih, Université de Montpellier, FR
Abstract
Since the expansion of the Internet of Things (IoT), connected devices became smart and autonomous. Their exponentially increasing number and their use in many application domains result in a huge potential of cybersecurity threats. Taking into account the evolution of the IoT, security and interoperability are the main challenges, to ensure the reliability of the information. The blockchain technology provides a new approach to handle the trust in a decentralized network. However, current blockchain implementations cannot be used in IoT domain because of their huge need of computing power and storage utilization. This demonstrator presents a lightweight distributed ledger protocol dedicated to the IoT application, reducing the computing power and storage utilization, handling the scalability and ensuring the reliability of information.
More information ...

UB02.9 PRE-IMPACT FALL DETECTION ARCHITECTURE BASED ON NEUROMUSCULAR CONNECTIVITY STATISTICS
Authors: Giovanni Mezzina, Sardar Mehboob Hussain and Daniela De Venuto, Politecnico di Bari, IT
Abstract
In this demonstration, we propose an innovative multi-sensor architecture operating in the field of pre-impact fall detection (PIFD). The proposed architecture jointly analyzes cortical and muscular involvement when unexpected slippages occur during steady walking. The EEG and EMG are acquired through wearable and wireless devices. The EMG computation block translates EMGs into binary signals, which are used both to enable cortical and muscular involvement when unexpected slippages occur during steady walking. The EEG and EMG are acquired through wearable and wireless devices.

More information ...

UB02.10 JOINTER: JOINING FLEXIBLE MONITORS WITH HETEROGENEOUS ARCHITECTURES
Authors: Giacomo Valente
1Universitá degli Studi dell’Aquila, IT; 2Università degli Studi di Sassari, IT; 3Università degli Studi di Cagliari, IT
Abstract
As embedded systems grow more complex and shift toward heterogeneous architectures, understanding workload performance characteristics becomes increasingly difficult. In this regard, run-time monitoring systems can support on obtaining the desired visibility to characterize a system. This demo presents a framework that allows to develop complex heterogeneous architectures composed of programmable processors and dedicated accelerators on FPGAs, together with customizable monitoring systems, keeping under control the introduced overhead. The whole development flow (and related prototyper EDA tools), that starts with the accelerators creation using a dataflow model, in parallel with the monitoring system customization using a library of elements, showing also the final joining, will be shown. Moreover, a comparison among different monitoring systems functionalities on different architectures developed on Zynq7000 SoC will be illustrated.
More information ...

15:00 End of session

3.0 LUNCHEON KEYNOTE SESSION
Date: Tuesday 10 March 2020
Time: 13:50 - 14:20
Location / Room: Amphithéâtre Jean Prouve
Chair: Marco Casale-Rossi, Synopsys, IT
Co-Chair:
Though neuromorphic systems were introduced decades ago, there has been a resurgence of interest in recent years due to the looming end of Moore’s law, the end of Dennard scaling, and the tremendous success of AI and deep learning for a wide variety of applications. With this renewed interest, there is a diverse set of research ongoing in neuromorphic computing, ranging from novel hardware implementations, device and materials to the development of new training and learning algorithms. There are many potential advantages to neuromorphic systems that make them attractive in today’s computing landscape, including the potential for very low power, efficient hardware that can perform neural network computation. Though some compelling results have been demonstrated thus far that demonstrate these advantages, there is still significant opportunity for innovations in hardware, algorithms, and applications in neuromorphic computing. In this talk, a brief overview of the history of neuromorphic computing will be discussed, and a summary of the current state of research in the field will be presented. Finally, a list of key challenges, open questions, and opportunities for future research in neuromorphic computing will be enumerated.
Accelerating Design Space Exploration efficiently is needed to optimize hardware accelerators. At high level, learning techniques can provide ways to either recognize previously synthesized kernels or to model the hidden dependences between synthesis directive costs and performances. At a lower level, speeding up RTL simulations based on data dependencies analysis can speed up one of the most time consuming steps.

This work proposes a method to accelerate the process of High-Level Synthesis (HLS) Design Space Exploration (DSE) by pre-characterizing micro-kernels offline and creating predictive models of these. HLS allows to generate different types of micro-architectures from the same untimed behavioral description. This is typically done by setting different combinations of synthesis options in the form or synthesis directives specified as pragmas in the code. This allows, e.g. to control how loops should be synthesized, arrays and functions. Unique combinations of these pragmas leads to micro-architectures with a unique area vs. performance/power trade-offs. The main problem is that the search space grows exponentially with the number of explorable operations. Thus, the main goal of efficient HLS DSE is to find the synthesis directives' combinations that lead to the Pareto-optimal designs quickly. Our proposed method is based on the pre-characterization of micro-kernels offline, creating predictive models for each of the kernels, and using the results to explore a new unseen behavioral description using compositional methods. In addition, we make use of perceptual hashing to match new unseen micro-kernels with the pre-characterized micro-kernels in order to further speed up the search process. Experimental results show that our proposed method is orders of magnitude faster than traditional methods.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

---

**PROSPECTOR: SYNTHESIZING EFFICIENT ACCELERATORS VIA STATISTICAL LEARNING**

**Abstract**

Accelerator design is expensive due to the effort required to understand an algorithm and optimize the design. Architects have embraced two technologies to reduce costs: High-level synthesis automatically generates hardware from code. Reconﬁgurable fabrics instantiate accelerators while avoiding fabrication costs for custom circuits. We further reduce design effort with statistical learning. We build an automated framework, called Prospector, that uses Bayesian techniques to optimize synthesis directives, reducing execution latency and resource usage in ﬁeld-programmable gate arrays. We show in a certain amount of time designs discovered by Prospector are closer to Pareto-optimal designs compared to prior approaches.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

---

**TANGO: AN OPTIMIZING COMPILER FOR JUST-IN-TIME RTL SIMULATION**

**Abstract**

With Moore's law coming to an end, the advent of hardware specialization presents a unique challenge for a much tighter software and hardware co-design environment to exploit domain-specific optimizations and increase design efficiency. This trend is further accentuated by rapid-pace of innovations in Machine Learning and Graph Analytic, calling for a faster product development cycle for hardware accelerators and the importance of addressing the increasing cost of hardware veriﬁcation. The productivity of software-hardware co-design relies upon a better integration between the software and hardware design methodologies, but more importantly in the effectiveness of the design tools and hardware simulators at reducing the development time. In this work, we developed Tango, an Optimizing compiler for a Just-In-Time RTL simulator. Tango implements unique hardware-centric compiler transformations to speed up runtime code generation in a software-hardware co-design environment where hardware simulation speed is critical. Tango achieves a 6x average speedup compared to the state-of-the-art RTL simulators.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

---

**POISONING THE (DATA) WELL IN ML-BASED CAD: A CASE STUDY OF HIDING LITHOGRAPHIC HOTSPOTS**

**Abstract**

Machine learning (ML) provides state-of-the-art performance in many parts of computer-aided design (CAD) flows. However, deep neural networks (DNNs) are susceptible to various adversarial attacks, including data poisoning to compromise training to insert backdoors. Sensitivity to training data integrity presents a security vulnerability, especially in light of malicious insiders who want to cause targeted neural network misbehavior. In this study, we explore this threat in lithographic hotspot detection via training data poisoning, where hotspots in a layout clip can be "hidden" at inference time by including a trigger shape in the input. We show that training data poisoning attacks are feasible and stealthy, demonstrating a backdoored neural network that performs normally on clean inputs but misbehaves on inputs when a backdoor trigger is present. Furthermore, our results raise some fundamental questions about the robustness of ML-based systems in CAD.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

---

**SOLOMON: AN AUTOMATED FRAMEWORK FOR DETECTING FAULT ATTACK VULNERABILITIES IN HARDWARE**

**Abstract**

Fault attacks are potent physical attacks on crypto-devices. A single fault injected during encryption can reveal the cipher's secret key. In a hardware realization of an encryption algorithm, only a tiny fraction of the gates is exploitable by such an attack. Finding these vulnerable gates has been a manual and tedious task requiring considerable expertise. In this paper, we propose SOLOMON, the first automatic fault attack vulnerability detection framework for hardware designs. Given a cipher implementation, either at RTL or gate-level, SOLOMON uses formal methods to map vulnerable regions in the cipher algorithm to specific locations in the hardware thus enabling targeted countermeasures to be deployed with much lesser overheads. We demonstrate the efficacy of the SOLOMON framework using three ciphers: AES, CLEFIA, and Simon.

**Download Paper (PDF; Only available from the DATE venue WiFi)**
**3.3 EU/ESA projects on Heterogeneous Computing**

**Date:** Tuesday 10 March 2020  
**Time:** 14:30 - 16:00  
**Location / Room:** Autrans

**Chair:**  
Carles Hernandez, UPV, ES

**Co-Chair:**  
Francisco J. Cazorla, BSC, ES

In the scope of this session the presented EU/ESA projects cover topics related to the control electronics and data processing architecture and functionality of the Wide Field Imager, one of two scientific instruments of the next European X-ray observatory ATHENA; task-based programming models to provide a software ecosystem for heterogeneous hardware composed of CPUs, GPUs, FPGAs and dataflow engines; and a framework to allow Big Data solutions to dynamically and transparently exploit heterogeneous hardware accelerators.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
</table>
| 15:00 | 3.3.2 | **LEGATO: LOW-ENERGY, SECURE, AND RESILIENT TOOLSET FOR HETEROGENEOUS COMPUTING**    | **Speaker:** Valerio Schiavoni, University of Neuchâtel, CH  
**Authors:** Behzad Salami1, Konstantinos Parasysris1, Adrian Cristal1, Osman Unsal1, Xavier Martorell1, Paul Carpenter1, Raul De La Cruz1, Leonardo Bautista1, Daniel Jimenez2, Carlos Alvarez2, Saber Nabavi2, Sergi Madonar2, Miquel Pericas2, Pedro Fracisco3, Mustafa Abduljabbar2, Jing Chen2, Pirah Noor Soomro2, Madhavan Manivannan2, Micha von dem Berge3, Stefan Krupop3, Frank Klawonn4, Amani Mikhlaifi2, Sigrun May1, Tobias Becker5, Georgi Gaydadjiev5, Hans Salomonson5, Devdutt Dubhashi5, Dron Port5, Yoav Etzion5, Le Quoc Do5, Christof Fetzer5, Martin Kaiser5, Nils Kucz5, Jens Hagemeyer5, René Griess5, Lennart Tiggeli5, Kevin Mika5, Arne Hüffmeier5, Marcelo Pasin11, Valerio Schiavoni11, Isabelly Rocha11, Christian Gütte11 and Pascal Felber11  
1BSC, ES;  
2Chalmers, SE;  
3Christmann Informationstechnik + Medien GmbH & Co. KG, DE;  
4Helmholtz-Zentrum für Infektionsforschung GmbH, DE;  
5MAXELER, GB;  
6MIS, SE;  
7TECHNION, IL;  
8Technion, IL;  
9TU Dresden, DE;  
10UNIBE, DE;  
11UNINE, CH

**Abstract**  
The LEGaTO project leverages task-based programming models to provide a software ecosystem for Made-in-Europe heterogeneous hardware composed of CPUs, GPUs, FPGAs and dataflow engines. The aim is to attain one order of magnitude energy savings from the edge to the converged cloud/HPC, balanced with the security and resilience challenges. LEGaTO is an ongoing three-year EU H2020 project started in December 2017.

Download Paper (PDF; Only available from the DATE venue WiFi)
EFFICIENT COMPILATION AND EXECUTION OF JVM-BASED DATA PROCESSING FRAMEWORKS ON HETEROGENEOUS CO-PROCESSORS

Authors:
Christos Kotselidis¹, Ioannis Komninos², Grestis Akrivopoulos², Sebastian Bress³, Katerina Doka⁴, Hazeef Mohammed⁵, Georgios Mylonas⁶, Vassilis Spitsadakis⁶, Ioannis Konstantinou7, Ioannis Mytilinos⁷, Constantinos Bitsakos⁷, Christos Tsalidis⁷, Christos Tselios⁷, Nikolaos Kanakis⁷, Clemens Lutz⁸, Viktor Rosenfeld⁹ and Volker Markl⁶

¹The University of Manchester, GB; ²Exeus Ltd., US; ³Spark Works ITC Ltd., GB; ⁴German Research Center for Artificial Intelligence, DE; ⁵National TU Athens, GR; ⁶Kaléo Ltd., GB; ⁷Computer Technology Institute & Press Diophantus, GR; ⁸Neurocom Luxembourg, LU; ⁹Proov Ltd., GB

Abstract
This paper addresses the fundamental question of how modern Big Data frameworks can dynamically and transparently exploit heterogeneous hardware accelerators. After presenting the major challenges that have to be addressed towards this goal, we describe our proposed architecture for automatic and transparent hardware acceleration of Big Data frameworks and applications. Our vision is to retain the uniform programming model of Big Data frameworks and enable automatic, dynamic Just-In-Time compilation of the candidate code segments that benefit from hardware acceleration to the corresponding format. In conjunction with machine learning-based device selection, that respect user-defined constraints (e.g., cost, time, etc.), we enable dynamic code execution on GPUs and FPGAs transparently to the user. In addition, we dynamically re-steer execution at runtime based on the availability of resources. Our preliminary results demonstrate that our approach can accelerate an existing Apache Flink application by up to 16.5x.

Download Paper (PDF; Only available from the DATE venue WiFi)

16:00 End of session

3.4 Accelerating Neural Networks and Vision Workloads

Date: Tuesday 10 March 2020
Time: 14:30 - 16:00
Location / Room: Stendhal

Chair: Leonidas Kosmidis, BSC, ES
Co-Chair: Georgios Keramidas, Aristotle University of Thessaloniki/Think Silicon S.A., GR

This session presents different solutions to accelerate emerging applications. The papers include various microarchitecture techniques as well as complete SoC and RISC-V based solutions. More fine-grained techniques are also presented like fast computations on sparse matrices. Vision applications are represented by the popular VSLAM, while various microcontrollers.

3.4.1 EFFICIENT COMPILATION AND EXECUTION OF JVM-BASED DATA PROCESSING FRAMEWORKS ON HETEROGENEOUS CO-PROCESSORS

Authors:
Athanasios Stratikopoulos, The University of Manchester, GB

Abstract
This paper addresses the fundamental question of how modern Big Data frameworks can dynamically and transparently exploit heterogeneous hardware accelerators. After presenting the major challenges that have to be addressed towards this goal, we describe our proposed architecture for automatic and transparent hardware acceleration of Big Data frameworks and applications. Our vision is to retain the uniform programming model of Big Data frameworks and enable automatic, dynamic Just-In-Time compilation of the candidate code segments that benefit from hardware acceleration to the corresponding format. In conjunction with machine learning-based device selection, that respect user-defined constraints (e.g., cost, time, etc.), we enable dynamic code execution on GPUs and FPGAs transparently to the user. In addition, we dynamically re-steer execution at runtime based on the availability of resources. Our preliminary results demonstrate that our approach can accelerate an existing Apache Flink application by up to 16.5x.

Download Paper (PDF; Only available from the DATE venue WiFi)

14:30 3.4.1 PSB-RNN: A PROCESSING-IN-MEMORY SYSTOLIC ARRAYARCHITECTURE USING BLOCK CIRCULANT MATRICES FOR RECURRENT NEURAL NETWORKS

Authors:
Nagadastigir Challapalle, Pennsylvania State University, US

Abstract
Recurrent Neural Networks (RNNs) are widely used in Natural Language Processing (NLP) applications as they inherently capture contextual information across spatial and temporal dimensions. Compared to other classes of neural networks, RNNs have more weight parameters as they primarily consist of fully connected layers. Recently, several techniques such as weight pruning, zero-skipping, and block circulant compression have been introduced to reduce the storage and access requirements of RNN weight parameters. In this work, we present a ReRAM crossbar based processing-in-memory (PIM) architecture with systolic dataflow incorporating block circulant compression for RNNs. The block circulant compression decomposes the operations in a fully connected layer into a series of Fourier transforms and point-wise operations resulting in reduced space and computational complexity. We formulate the Fourier transform and point-wise operations into in-situ multiply-and-accumulate (MAC) operations mapped to ReRAM crossbars for high energy efficiency and throughput. We also incorporate systolic dataflow for communication within the crossbar arrays, in contrast to broadcast and multicast communications, to further improve energy efficiency. The proposed architecture achieves average improvements in compute efficiency of 44x and 17x over a custom FPGA and conventional crossbar based architecture implementations, respectively.

Download Paper (PDF; Only available from the DATE venue WiFi)

15:00 3.4.2 XPULPN: ACCELERATING QUANTIZED NEURAL NETWORKS ON RISC-V PROCESSORS THROUGH ISA EXTENSIONS

Authors:
Angelo Garofalo, Università di Bologna, IT

Abstract
Strongly quantized fixed-point arithmetic is considered the key direction to enable the inference of CNNs on low-power, resource-constrained edge devices. However, the deployment of highly quantized Neural Networks at the extreme edge of IoT, on fully programmable MCUs, is currently limited by the lack of support, at the Instruction Set Architecture (ISA) level, for sub-byte fixed-point data types, making it necessary to add numerous instructions for packing and unpacking data when running low-bitwidth (i.e. 2- and 4-bit) QNN kernels, creating a bottleneck for performance and energy efficiency of QNN inference. In this work we present a set of extensions to the RISC-V ISA, aimed at boosting the energy efficiency of low-bitwidth QNNs on low-power microcontroller-class cores. The microarchitecture supporting the new extensions is built on top of a RISC-V core featuring instruction set extensions targeting energy-efficient digital signal processing. To evaluate the extensions, we integrated the core into a full microcontroller system, synthesized and placed/routed in 22nm FDX technology. QNN convolution kernels, implemented on the new core, run 5.3x and 8.9x faster when considering 4- and 2-bit data operands respectively, compared to the baseline processor only supporting 8-bit SIMD instructions. With a peak of 279 GMAC/s/W, the proposed solution achieves 9x better energy efficiency compared to the baseline and two orders of magnitudes better energy efficiency compared to state-of-the-art microcontrollers.

Download Paper (PDF; Only available from the DATE venue WiFi)
3.5 Parallel real-time systems

Date: Tuesday 10 March 2020
Time: 14:30 - 16:00
Location / Room: Bayard
Chair: Liliana Cucu-Grosjean, Inria, FR
Co-Chair: Antoine Bertout, ENSMA, FR

This session presents novel techniques to enable parallel execution in real-time systems. More precisely, the papers are solving limitations of previous DAG models, devising tool chains to ensure WCET bounds, correcting results on heterogeneous processors, and considering wireless networks with application-oriented scheduling.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:30</td>
<td>3.4.3</td>
<td><strong>SNA: A SIAMESE NETWORK ACCELERATOR TO EXPLOIT THE MODEL-LEVEL PARALLELISM OF HYBRID NETWORK STRUCTURE</strong></td>
<td>Xingbin Wang, Chinese Academy of Sciences, CN</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Speaker:</strong> Xingbin Wang, Boney Zhao, Rui Hou and Dan Meng, Chinese Academy of Sciences, CN</td>
<td><strong>Authors:</strong> Xingbin Wang, Boney Zhao, Rui Hou and Dan Meng, Chinese Academy of Sciences, CN</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Siamese network is a compute-intensive learning model with growing applicability in a wide range of domains. However, state-of-art deep neural network (DNN) accelerators would not work efficiently for siamese network, as their designs do not account for the algorithm properties of siamese network. In this paper, we propose a siamese network accelerator called SNA, the first Simultaneous Multi-Threading (SMT) hardware architecture to perform siamese network inference with high performance and energy efficiency. We devise an adaptive inter-model computing resource partition and flexible on-chip buffer management mechanism based on the model parallelism and SMT design philosophy. Our architecture is implemented in Verilog and synthesized in a 65nm technology using Synopsys design tools. We also evaluate it with several typical siamese networks. Compared to the state-of-art accelerator, on average, the SNA architecture offers 2.1x speedup and 1.48x energy reduction.</td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
<tr>
<td>15:45</td>
<td>3.4.4</td>
<td><strong>HCVEACC: A HIGH-PERFORMANCE AND ENERGY-EFFICIENT ACCELERATOR FOR TRACKING TASK IN VSLAM SYSTEM</strong></td>
<td>Li Renwei, Wu Junning, Liu Meng, Chen Zuding, Zhou Shengang and Feng Shanggong, Chinese Academy of Sciences, CN</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Speaker:</strong> Meng Liu, Chinese Academy of Sciences, CN</td>
<td><strong>Authors:</strong> Meng Liu, Chinese Academy of Sciences, CN</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Visual SLAM (vSLAM) is a critical computer vision technology that is able to build a map of an unknown environment and perform location, simultaneously leveraging the partially built map. While existing several software SLAM processing frameworks, underlying general-purpose processors still hardly achieve the real-time SLAM at a reasonably low cost. In this paper, we propose HcveAcc, the first specialized CMOS-based hardware accelerator to help optimize the tracking task in the vSLAM system with high-performance and energy-efficient. Our HcveAcc targets to solve the time overhead bottleneck in the tracking process—high-density feature extraction and high-precision descriptor generation, and provides a configurable hardware architecture that handles higher resolution image data. We have implemented the HcveAcc in a 28nm CMOS technology using commercial EDA tools and evaluated it for the EuRoC and TUM dataset to demonstrate the robustness and accuracy in the SLAM tracking procedure. Our results show that HcveAcc achieves 4.3X speedup while consuming much less energy compared with state-of-the-art FPGA solutions.</td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
<tr>
<td>16:01</td>
<td>645</td>
<td><strong>ACCELERA: ACCELERATING SPARSE COMPUTATION BY ENABLING STREAM ACCESSES TO MEMORY</strong></td>
<td>Bahar Asgari, Georgia Tech, US</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Speaker:</strong> Bahar Asgari, Georgia Tech, US</td>
<td><strong>Authors:</strong> Bahar Asgari, Ramyad Hadidi and Hyesoon Kim, Georgia Tech, US</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Sparse computations dominate a wide range of applications from scientific problems to graph analytics. The main characterization of sparse computations, indirect memory accesses, prevents them from effectively achieving high performance on general-purpose processors. Therefore, hardware accelerators have been proposed for sparse problems. For these accelerators, the storage format and the decompression mechanism is crucial but have seen less attention in prior work. To address this gap, we propose Ascella, an accelerator for sparse computations, which besides enabling a smooth stream of data and parallel computation, proposes a fast decompression mechanism. Our implementation on a ZYNQ FPGA shows that on average, Ascella executes sparse problems up to 5.1x as fast as prior work.</td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
<tr>
<td>16:02</td>
<td>645</td>
<td><strong>ACCELERATION OF PROBABILISTIC REASONING THROUGH CUSTOM PROCESSOR ARCHITECTURE</strong></td>
<td>Nimish Shah, KU Leuven, BE</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Speaker:</strong> Nimish Shah, Laura I. Galindez Olassoaga, Wannes Meert and Marian Verhelst, KU Leuven, BE</td>
<td><strong>Authors:</strong> Nimish Shah, Laura I. Galindez Olassoaga, Wannes Meert and Marian Verhelst, KU Leuven, BE</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Probabilistic reasoning is an essential tool for robust decision-making systems because of its ability to explicitly handle real-world uncertainty, constraints and causal relations. Consequently, researchers are developing hybrid models by combining Deep Learning with Probabilistic reasoning for safety-critical applications like self-driving vehicles, autonomous drones, etc. However, probabilistic reasoning kernels do not execute efficiently on CPUs or GPUs. This paper, therefore, proposes a custom programmable processor to accelerate sum-product networks, an important probabilistic reasoning execution kernel. The processor has an optimized datapath architecture and memory hierarchy optimized for sum-product networks execution. Experimental results show that the processor, while requiring fewer computational and memory units, achieves a 12x throughput benefit over the Nvidia Jetson TX2 embedded GPU platform.</td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
</tbody>
</table>

16:00 End of session
ON THE VOLUME CALCULATION FOR CONDITIONAL DAG TASKS: HARDNESS AND ALGORITHMS

Abstract

The hardness of analyzing conditional directed acyclic graph (DAG) tasks remains unknown so far. For example, previous researches asserted that the conditional DAG's volume can be solved in polynomial time. However, these researches all assume well-nested structures that are recursively composed by single-source-single-sink parallel and conditional components. For conditional DAGs in general that do not comply with this assumption, the hardness and algorithms of volume computation are still open. In this paper, we construct counterexamples to show that previous work cannot provide a safe upper bound of the conditional DAG's volume in general. Moreover, we prove that the volume computation problem for conditional DAGs is strongly NP-hard. Finally, we propose an exact algorithm for computing the conditional DAG's volume. Experiments show that our method can significantly improve the accuracy of the conditional DAG's volume estimation.

Download Paper (PDF; Only available from the DATE venue WiFi)

WCET-AWARE CODE GENERATION AND COMMUNICATION OPTIMIZATION FOR PARALLELINGING COMPILERS

Abstract

High performance demands of present and future embedded applications increase the need for multi-core processors in hard real-time systems. Challenges in static multi-core WCET-analysis and the more complex design of parallel software, however, oppose the adoption of multi-core processors in that area. Automated parallelization is a promising approach to solve these issues, but specialized solutions are required to preserve static analyzability. With a WCET-aware parallelizing transformation, this work presents a novel solution for an important building block of a real-time capable parallelizing compiler. The approach includes a technique to optimize communication and synchronization in the parallelized program and supports complex memory hierarchies consisting of shared and core-private memory segments. In an experiment with four different applications, the parallelization improved the WCET by up to factor 3.2 on 4 cores. The studied optimization technique and the support for shared memories significantly contribute to these results.

Download Paper (PDF; Only available from the DATE venue WiFi)

TEMPLATE SCHEDULE CONSTRUCTION FOR GLOBAL REAL-TIME SCHEDULING ON UNRELATED MULTIPROCESSOR PLATFORMS

Abstract

The seminal work on the global real-time scheduling of periodic tasks on unrelated multiprocessor platforms is based on a two-steps method. First, the workload of each task is distributed over the processors and it is proved that the success of this first step ensures the existence of a feasible schedule. Second, a method for the construction of a template schedule from the workload assignment is presented. In this work, we review the seminal work and show by using a counter-example that this second step is incomplete. Thus, we propose and prove correct a novel and efficient algorithm to build the template schedule.

Download Paper (PDF; Only available from the DATE venue WiFi)

APPLICATION-AWARE SCHEDULING OF NETWORKED APPLICATIONS OVER THE LOW-POWER WIRELESS BUS

Abstract

Recent successes of wireless networked systems in advancing industrial automation and in spawning the Internet of Things paradigm motivate the adoption of wireless networked systems in current and future safety-critical applications. As reliability is key in safety-critical applications, in this work we present NetDAG, a scheduler design and implementation suitable for real-time applications in the wireless setting. NetDAG is built upon the Low-Power Wireless Bus, a high-performant communication abstraction for wireless networked systems, and enables system designers to directly schedule applications under specified task-level real-time constraints. Access to real-time primitives in the scheduler permits efficient design exploration of tradeoffs between power consumption and latency. Furthermore, NetDAG provides support for weakly hard real-time applications with deterministic guarantees, in addition to heretofore considered soft real-time applications with probabilistic guarantees. We propose novel abstraction techniques for reasoning about conjunctions of weakly hard constraints and show how such abstractions can be used to handle the significant scheduling difficulties brought on by networked components with weakly hard behaviors.

Download Paper (PDF; Only available from the DATE venue WiFi)

A PERFORMANCE ANALYSIS FRAMEWORK FOR REAL-TIME SYSTEMS SHARING MULTIPLE RESOURCES

Abstract

Timing properties of applications strongly depend on resources that are allocated to them. Applications often have multiple resource requirements, all of which must be met for them to proceed. Performance analysis of event-based systems has been widely studied in the literature. However, the proposed works consider only one resource requirement for each application task. Additionally, they mainly focus on the rate at which resources serve applications (e.g., power, instructions or bits per second), but another aspect of resources, which is their provided capacity (e.g., energy, memory ranges, FPGA regions), has been ignored. In this work, we propose a mathematical framework to describe the provisioning rate and capacity of various types of resource. Additionally, we consider the simultaneous use of multiple resources. Conservative bounds on response times of events and their backlog are computed. We prove that the bounds are monotone in event arrivals and in required and provided rate and capacity, which enables verification of real-time application performance based on worst-case characterizations. The applicability of our framework is shown in a case study.

Download Paper (PDF; Only available from the DATE venue WiFi)
and new technologies, such as ReRAM.

Based architectures must respond to highly demanding applications such as image segmentation and classification by taking advantage of new topologies (multiple layers, 3D...). To support innovative applications, new paradigms have been introduced, such as neural network and approximate computing. This session presents different NoC-based computing system. Different mechanisms are introduced at network-level to support the application and thus enhance the performance and power efficiency. As such, new NoC-based architectures that support these computing approaches.

To support innovative applications, new paradigms have been introduced, such as neural network and approximate computing. This session presents different NoC-based architectures that support these computing approaches. In these advanced architectures, NoC designs are no longer only a communication infrastructure but also part of the computing system. Different mechanisms are introduced at network-level to support the application and thus enhance the performance and power efficiency. As such, new NoC-based architectures must respond to highly demanding applications such as image segmentation and classification by taking advantage of new topologies (multiple layers, 3D...) and new technologies, such as ReRAM.

### 3.6 NoC in the age of neural network and approximate computing

**Date:** Tuesday 10 March 2020  
**Time:** 14:30 - 16:00  
**Location/Room:** Lesdiguières

**Chair:**  
Romain Lemaire, CEA-Leti, FR

To support innovative applications, new paradigms have been introduced, such as neural network and approximate computing. This session presents different NoC-based architectures that support these computing approaches. In these advanced architectures, NoC designs are no longer only a communication infrastructure but also part of the computing system. Different mechanisms are introduced at network-level to support the application and thus enhance the performance and power efficiency. As such, new NoC-based architectures must respond to highly demanding applications such as image segmentation and classification by taking advantage of new topologies (multiple layers, 3D...) and new technologies, such as ReRAM.

**Authors**

Matheus Schuh, Verimag / Kalray, FR  
Maximilien Dupont de Dinechin, Matthieu Moy and Claire Maiza

**Abstract**

In RTNS 2016, Riani et al. proposed an algorithm to compute the impact of interference on memory accesses on the timing of a task graph. It calculates a static, time-triggered schedule, i.e., a release date and a worst-case response time for each task. The task graph is a DAG, typically obtained by compilation of a high-level dataflow language, and the tool assumes a previously determined mapping and execution order. The algorithm is precise, but suffers from a high O(n^2) complexity, n being the number of input tasks. Since we target many-core platforms with tens or hundreds of cores, applications likely to exploit the parallelism of these platforms are too large to be handled by this algorithm in reasonable time. Instead of performing global fixed-point iterations on the task graph, we compute the static schedule incrementally, reducing the complexity to O(n^2). Experimental results show a reduction from 535 seconds to 0.90 seconds on a benchmark with 384 tasks, i.e. 593 times faster.

Download Paper (PDF; Only available from the DATE venue WiFi)
LIGHTWEIGHT ANONYMOUS ROUTING IN NOC BASED SOCS
Speaker: Prabhat Mishra, University of Florida, US
Authors: Subodha Charles, Megan Logan and Prabhat Mishra, University of Florida, US
Abstract
System-on-Chip (SoC) supply chain is widely acknowledged as a major source of security vulnerabilities. Potentially malicious third-party IPs integrated on the same Network-on-Chip (NoC) with the trusted components can lead to security and trust concerns. While secure communication is a well studied problem in computer networks domain, it is not feasible to implement those solutions on resource-constrained SoCs. In this paper, we present a lightweight anonymous routing protocol for communication between IP cores in NoC based SoCs. Our method eliminates the major overhead associated with traditional anonymous routing protocols while ensuring that the desired security goals are met. Experimental results demonstrate that existing security solutions on NoC can introduce significant (1.5X) performance degradation, whereas our approach provides the same security features with minor (4%) impact on performance.

Download Paper (PDF; Only available from the DATE venue WiFi)

3.7 Augmented and Assisted Living: A reality
Date: Tuesday 10 March 2020
Time: 14:30 - 16:00
Location / Room: Berlioz
Chair: Graziano Pravadelli, Università di Verona, IT
Co-Chair: Vassilis Pavlidis, Aristotle University of Thessaloniki, GR
Novel solutions for healthcare and ambient assistant living: innovative brain-computer interfaces, novel cancer prediction systems and energy-efficient ECG and wearable systems.

14:30 3.7.1 COMPRESSION SUBJECT-SPECIFIC BRAIN-COMPUTER INTERFACE MODELS INTO ONE MODEL BY SUPERPOSITION IN HYPERDIMENSIONAL SPACE
Speaker: Michael Hersche, ETH Zurich, CH
Authors: Michael Hersche, Philipp Rupp, Luca Benini and Abbas Rahimi, ETH Zurich, CH
Abstract
Accurate multiclass classification of electroencephalography (EEG) signals is still a challenging task towards the development of reliable motor imagery brain-computer interfaces (MI-BCIs). Deep learning algorithms have been recently used in this area to deliver a compact and accurate model. Reaching high-level of accuracy requires to store subjects-specific trained models that cannot be achieved with an otherwise compact model trained globally across all subjects. In this paper, we propose a new methodology that closes the gap between these two extreme modeling approaches: we reduce the overall storage requirements by superimposing many subject-specific models into one single model such that it can be reliably decomposed, after retraining, to its constituent models while providing a trade-off between compression ratio and accuracy. Our method makes the use of unexploited capacity of trained models by orthogonalizing parameters in a hyperdimensional space, followed by iterative retraining to compensate noisy decomposition. This method can be applied to various layers of deep inference models. Experimental results on the 4-class BCI competition IV-2a dataset show that our method exploits unutilized capacity for compression and surpasses the accuracy of two state-of-the-art networks: (1) it compresses the smallest network, EEGNet [1], by 1.9x, and increases its accuracy by 2.41% (74.73% vs. 72.32%); (2) using a relatively larger Shallow ConvNet [2], our method achieves 2.95x compression as well as 1.4% higher accuracy (75.05% vs. 73.59%).

Download Paper (PDF; Only available from the DATE venue WiFi)

15:00 3.7.2 A NOVEL FPGA-BASED SYSTEM FOR TUMOR GROWTH PREDICTION
Speaker: Yannis Papaefstathiou, Aristotle University of Thessaloniki, GR
Authors: Konstantinos Malavazos 1, Maria Papadogiorgaki 1, PAVLOS MALAKONAKIS 1 and Ioannis Papaefstathiou 2
1TU Crete, GR; 2Aristotle University of Thessaloniki, GR
Abstract
An emerging trend in the biomedical community is to create models that take advantage of the increasing available computational power, in order to manage and analyze new biological data as well as to model complex biological processes. Such biomedical software applications require significant computational resources since they process and analyze large amounts of data, such as medical image sequences. This paper presents a novel FPGA-based system that implements a novel model for the prediction of the spatio-temporal evolution of glioma. Glioma is a rapidly evolving type of brain cancer, well known for its aggressive and diffusive behavior. The developed system simulates the glioma tumor growth in the brain tissue, which consists of different anatomic structures, by utilizing individual MRI slices. The presented innovative hardware system is more than 60% faster than a high-end server consisting of 20 physical cores (and 40 virtual ones) and more than 28x more energy efficient.

Download Paper (PDF; Only available from the DATE venue WiFi)
15:30 3.7.3 AN EVENT-BASED SYSTEM FOR LOW-POWER ECG QRS COMPLEX DETECTION

**Presentation Title:** AN EVENT-BASED SYSTEM FOR LOW-POWER ECG QRS COMPLEX DETECTION

**Authors:** Silvio Zanoli, EPFL, CH

**Abstract**

One of the greatest challenges in the design of modern wearable devices is energy efficiency. While data processing and communication have received a lot of attention from the industry and academia, leading to highly efficient microcontrollers and transmission devices, sensor data acquisition in medical devices is still based on a conservative paradigm that requires regular sampling at the Nyquist rate of the target signal. This requirement is usually excessive for sparse and highly non-stationary signals, leading to data overload and a waste of resources in the full processing pipeline. In this work, we propose a new system to create event-based heart-rate analysis devices, including a novel algorithm for QRS detection that is able to process electrocardiogram signals acquired irregularly and much below the theoretically-required Nyquist rate. This technique allows us to drastically reduce the average sampling frequency of the signal and, hence, the energy needed to process it and extract the relevant information. We implemented both the proposed event-based algorithm and a state-of-the-art version based on regular Nyquist rate based sampling on an ultra-low power hardware platform, and the experimental results show that the event-based version reduces the energy consumption in runtime up to 15.6 times, while the detection performance is maintained at an average F1 score of 99.5%.

[Download Paper (PDF; Only available from the DATE venue WiFi)]

15:45 3.7.4 SEMI-AUTONOMOUS PERSONAL CARE ROBOTS INTERFACE DRIVEN BY EEG SIGNALS DIGITIZATION

**Presentation Title:** SEMI-AUTONOMOUS PERSONAL CARE ROBOTS INTERFACE DRIVEN BY EEG SIGNALS DIGITIZATION

**Authors:** Daniela De Venuto, Politecnico di Bari, IT

**Abstract**

In this paper, we propose an innovative architecture that merges the Personal Care Robots (PCRs) advantages with a novel Brain Computer Interface (BCI) to carry out assistive tasks, aiming to reduce the burdens of caregivers. The BCI is based on movement related potentials (MRPs) and exploits EEG from 8 smart wireless electrodes placed on the sensorimotor area. The collected data are firstly pre-processed and then sent to a novel Feature Extraction (FE) step. The FE stage is based on symbolization algorithm, the Local Binary Patternning, which adopts end-to-end binary operations. It strongly reduces the stage complexity, speeding the BCI up. The final user intentions discrimination is entrusted to a linear Support Vector Machine (SVM). The BCI performances have been evaluated on four healthy young subjects. Experimental results showed a user intention recognition accuracy of ~84% with a timing of ~554 ms per decision. A proof of concept is presented, showing how the BCI-based binary decisions could be used to drive the PCR up to a requested object, expressing the will to keep it (delivering it to user) or to continue the research.

[Download Paper (PDF; Only available from the DATE venue WiFi)]

16:01 18, IP1-

**Presentation Title:** A NON-INVASIVE WEARABLE BIOIMPEDANCE SYSTEM TO WIRELESSLY MONITOR BLADDER FILLING

**Authors:** Markus Reichmuth, Simone Schuerle and Michele Magno, ETH Zurich, CH

**Abstract**

Monitoring of renal function can be crucial for patients in acute care settings. Commonly during postsurgical surveillance, urinary catheters are employed to assess the urine output accurately. However, as any external device inserted into the body, the use of these catheters carries a significant risk of infection. In this paper, we present a non-invasive method to measure the fill rate of the bladder, and thus the rate of renal clearance, via an external bioimpedance sensor system to avoid the use of urinary catheters, thereby eliminating the risk of infections and improving patient comfort. We design and propose a 4-electrode front-end and the whole wearable and wireless system with low power and accuracy in mind. The results demonstrate the accuracy of the sensors and low power consumption of only 80µW with a duty cycling of 1 acquisition every 5 minutes, which makes this battery-operated wearable device a long-term monitor system.

[Download Paper (PDF; Only available from the DATE venue WiFi)]

16:02 19, IP1-

**Presentation Title:** INFINIWOLF: ENERGY EFFICIENT SMART BRACELET FOR EDGE COMPUTING WITH DUAL SOURCE ENERGY HARVESTING

**Authors:** Michele Magno, ETH Zurich, CH

**Abstract**

This work presents InfiniWolf, a novel multi-sensor smartwatch that can achieve self-sustainability exploiting thermal and solar energy harvesting, performing computationally high demanding tasks. The smartwatch embeds both a System-on-Chip (SoC) with an ARM Cortex-M processor and Bluetooth Low Energy (BLE) and Mr. Wolf, an open-hardware RISC-V based parallel ultra-low-power processor that boosts the processing capabilities on board by more than one order of magnitude, while also increasing energy efficiency. We demonstrate its functionality based on a sample application scenario performing stress detection with multi-layer artificial neural networks on a wearable multi-sensor bracelet. Experimental results show the benefits in terms of energy efficiency and latency of Mr. Wolf over an ARM Cortex-MAF micro-controllers and the possibility, under specific assumptions, to be self-sustainable using thermal and solar energy harvesting while performing up to 24 stress classifications per minute in indoor conditions.

[Download Paper (PDF; Only available from the DATE venue WiFi)]

16:00

End of session
FLETCHER: TRANSPARENT GENERATION OF HARDWARE INTERFACES FOR ACCELERATING BIG DATA APPLICATIONS

Authors: Zaid Al-Ars, Johan Peルtenburg, Jeroen van Straten, Matthijs Brobbel and Joost Hoozemans, TU Delft, NL

Abstract
This demo created by TUDelft is a software-hardware framework to allow for an efficient integration of FPGA hardware accelerators both on edge devices as well as in the cloud. The framework is called Fletcher, which is used to automatically generate data communication interfaces in hardware based on the widely used big data format Apache Arrow. This provides two distinct advantages. On the one hand, since the accelerators use the same data format as the software, data communication bottlenecks can be reduced. On the other hand, since a standardized data format is used, this allows for easy-to-use interfaces on the accelerator side, thereby reducing the design and development time. The demo shows how to use Fletcher for big data acceleration to decompresse Snappy compressed files and perform filtering on the whole Wikipedia body of text. The demo enables 25 GB/s processing throughput.

More information ...

ELSA: EIGENVALUE BASED HYBRID LINEAR SYSTEM ABSTRACTION: BEHAVIORAL MODELING OF TRANSISTOR-LEVEL CIRCUITS USING AUTOMATIC ABSTRACTION TO HYBRID AUTOMATA

Authors: Ahmad Tarraf and Lars Hedrich, University of Frankfurt, DE

Abstract
Model abstraction of transistor-level circuits, while preserving an accurate behavior, is still an open problem. In this demo an approach is presented that automatically generates a hybrid automaton (HA) with linear states from an existing circuit netlist. The approach starts with a netlist at transistor level with full SPICE accuracy and ends at the system level description of the circuit in Matlab or in Verilog-A. The resulting hybrid automaton exhibits linear behavior as well as the accurate and speed-up of the Verilog-A generated models is evaluated based on several transistor level circuit abstractions of simple operational amplifiers up to a complex filters. Moreover, we verify the equivalence between the generated model and the original circuit. For the generated models in matlab syntax, a reachability analysis is performed using the reachability tool cara.

More information ...

PAFUSI: PARTICLE FILTER FUSION ASIC FOR INDOOR POSITIONING

Authors: Christian Schott, Marko Rößler, Daniel Froß, Marcel Putsche and Ulrich HeinkeI, TU Chemnitz, DE

Abstract
The meaning of data acquired from IoT devices is heavily enhanced if global or local position information of their acquisition is known. Infrastructure for indoor positioning as well as the IoT device involves heavy small, energy efficient but powerful devices that provide the location awareness. We propose the PAFUSI, a hardware implementation of an UWB position estimation algorithm that fulfills these requirements. Our design fuses distance measurements to fixed points in an environment to calculate the position in 3D space and is capable of using different positioning techniques like SPS, DeadWave or Nanotro as data source simultaneously. Our design comprises of an estimator which processes the data by means of a Sequential Monte Carlo method and a microcontroller core which configures and controls the measurement unit as well as it analyses the results of the estimator. The PAFUSI is manufactured as a monolithic integrated ASIC in a multi-project wafer in UMC's 65nm process.

More information ...

FPGA-DSP: A PROTOTYPE FOR HIGH QUALITY DIGITAL AUDIO SIGNAL PROCESSING BASED ON AN FPGA

Authors: Bernhard Riess and Christian Epe, University of Applied Sciences Düsseldorf, DE

Abstract
Our demonstrator presents a prototype of a new digital audio signal processing system which is based on an FPGA. It achieves a performance that up to now has been preserved to costly high-end solutions. Main components of the system are an analog/digital converter, an FPGA to perform the digital signal processing tasks, and a digital/analog converter implemented on a printed circuit board. To demonstrate the quality of the audio signal processing, infinite impulse response, finite impulse response filters and a delay effect were realized in VHDL. More advanced signal processing systems can easily be implemented due to the flexibility of the FPGA. Measured results were compared to state of the art audio signal processing systems with respect to size, performance and cost. Our prototype outperforms systems of the same price in quality, and outperforms systems of the same quality at a maximum of 20% of the price. Examples of the performance of our system can be heard in the demo.

More information ...

LEARNV: LEARNV: A RISC-V BASED EMBEDDED SYSTEM DESIGN FRAMEWORK FOR EDUCATION AND RESEARCH DEVELOPMENT

Authors: Noureddine Ait Said and Mourin Benabdenbi, TIMA Laboratory, FR

Abstract
Designing a modern System on a Chip is based on the joint design of hardware and software (co-design). However, understanding the tight relationship between hardware and software is not straightforward. Moreover to validate new concepts in SoC design from the idea to the hardware implementation is time-consuming and often slowed by legacy issues (intellectual property of hardware blocks and expensive commercial tools). To overcome these issues we propose to use the open-source Rocket Chip environment for educational purposes, combined with the open-source LowRisc architecture to implement a custom SoC design on an FPGA board. The demonstration will present how students and engineers can take benefit from the environment to deepen their knowledge in HW and SW co-design. Using the LowRisc architecture, an image classification application based on the use of CNNs will serve as a demonstrator of the whole open-source hardware and software flow and will be mapped on a Nexys A7 FPGA board.

More information ...

CSI-REPUTE: A LOW POWER EMBEDDED DEVICE CLUSTERING APPROACH TO GENOME READ MAPPING

Authors: TouSif Rahman1, Siddarth Maheshwari1, Rishad Shafik2, Ian Wilson1, Alex Yakovlev1 and Amit Acharyya2
1Newcastle University, GB; 2IIT Hyderabad, IN

Abstract
The big data challenge of genomics is rooted in its requirements of extensive computational capability and results in large power and energy consumption. To encourage widespread usage of genome assembly tools there must be a transition from the existing predominantly software-based mapping tools, optimized for homogeneous high-performance systems, to more heterogeneous low power and cost-effective mapping systems. This demonstration will show a cluster system implementation for the REPUTE algorithm, (An OpenCL based Read Mapping Tool for Embedded Genomics) where cluster nodes are composed of low power single board computer (SBC) devices and the algorithm is deployed on each node spreading the genomic workload. We propose a working concept prototype to challenge the technology dependent nonlinear e.g. limiting behavior. The accuracy and speed-up of the Verilog-A generated models is evaluated based on several transistor level circuit abstractions of simple operational amplifiers up to a complex filters. Moreover, we verify the equivalence between the generated model and the original circuit. For the generated models in matlab syntax, a reachability analysis is performed using the reachability tool cara.

More information ...

FUZZING EMBEDDED BINARIES LEVERAGING SYSTEMC-BASED VIRTUAL PROTOTYPES

Authors: Vladimir Herdt1, Daniel Grosse2 and Rolf Drechsler2
1DFKI, DE; 2University of Bremen / DFKI GmbH, DE

Abstract
Verification of embedded Software (SW) binaries is very important. Mainly, simulation-based methods are employed that execute (randomly) generated test-cases on Virtual Prototypes (VPs). However, to enable a comprehensive VP-based verification, sophisticated test-case generation techniques need to be integrated. Our demonstrator combines state-of-the-art fuzzing techniques with SystemC-based VPs to enable a fast and accurate verification of embedded SW binaries. The fuzzing process is guided by the coverage of the embedded SW as well as the SystemC-based peripherals of the VP. The effectiveness of our approach is demonstrated by our experiments, using RISC-V SW binaries as an example.

More information ...
**WALLANCE: AN ALTERNATIVE TO BLOCKCHAIN FOR IOT**
Authors: Loïc Dalmasso, Florent Bruguer, Pascal Benoit and Achraf Lamlih, Université de Montpellier, FR

**Abstract**
Since the expansion of the Internet of Things (IoT), connected devices became smart and autonomous. Their exponentially increasing number and their use in many application domains result in a huge potential of cybersecurity threats. Taking into account the evolution of the IoT, security and interoperability are the main challenges, to ensure the reliability of the information. The blockchain technology provides a new approach to handle the trust in a decentralized network. However, current blockchain implementations cannot be used in IoT domain because of their huge need of computing power and storage utilization. This demonstrator presents a lightweight distributed ledger protocol dedicated to the IoT application, reducing the computing power and storage utilization, handling the scalability and ensuring the reliability of information.

**More information ...**

**RUMORE: A FRAMEWORK FOR RUNTIME MONITORING AND TRACE ANALYSIS FOR COMPONENT-BASED EMBEDDED SYSTEMS DESIGN FLOW**
Authors: Vittoriano Mutilillo1, Luigi Pomante1, Giacomo Valente1, Hector Posadas2, Javier Merino2 and Eugenio Villar2
1University of L'Aquila, IT; 2University of Cantabria, ES

**Abstract**
The purpose of this demonstrator is to introduce runtime monitoring infrastructures and to analyze trace data. The goal is to show the concept among different monitoring requirements by defining a general reference architecture that can be adapted to different scenarios. Starting from design artifacts, generated by a system engineering modeling tool, a custom HW monitoring system infrastructure will be presented. This sub-system will be able to generate runtime artifacts for runtime verification. We will show how the RUMORE framework provides round-trip support in the development chain, injecting monitoring requirements from design models down to code and its execution on the platform and trace data back to the models, where the expected behavior will then compared with the actual behavior.

This approach will be used towards optimizing design models for specific properties (e.g. for system performance).

**More information ...**

**FASTHERMSIM: FAST AND ACCURATE THERMAL SIMULATIONS FROM CHIPLETS TO SYSTEM**
Authors: Yu-Min Lee, Chi-Wen Pan, Li-Rui Ho and Hong-Wen Chiou, National Chiao Tung University, TW

**Abstract**
Recently, owing to the scaling down of technology and 2.5D/3D integration, power densities and temperatures of chips have been increasing significantly. Though commercial computational fluid dynamics tools can provide accurate thermal maps, they may lead to inefficiency in thermal-aware design with huge runtime. Thus, we develop the chip/package/system-level thermal analyzer, called FastThermSim, which can assist you to improve your design under thermal constraints in pre/post-silicon stages. In FastThermSim, we consider three heat transfer modes, conduction, convection, and thermal radiation. We convert them to temperature-independent terms by linearization methods and build a compact thermal model (CTM). By applying numerical methods to the CTM, the steady-state and transient thermal profiles can be solved efficiently without loss of accuracy. Finally, an easy-to-use thermal analysis tool is implemented for your design, which is flexible and compatible, with the graphic user interface.

**More information ...**

**DYNUNLOCK: UNLOCKING SCAN CHAINS OBfuscated USING DYNAMIC KEYS**
Speaker: Nimisha Limaye, New York University, US
Authors: Nimisha Limaye1 and Ozgur Sinanoglu 2
1New York University, US; 2New York University Abu Dhabi, AE

**Abstract**
Outsourcing in semiconductor industry opened up venues for faster and cost-effective chip manufacturing. However, this also introduced untrusted entities with malicious intent, to steal intellectual property (IP), overproduce the circuits, insert hardware Trojans, or counterfeit the chips. Recently, a defense is proposed to obfuscate the scan access based on a dynamic key that is initially generated from a secret key but changes in every clock cycle. This defense can be considered as the most rigorous defense among all the scan locking techniques. In this paper, we propose an attack that remolds this defense into one that can be broken by the SAT attack, while we also note that our attack can be adjusted to break other less rigorous (key that is updated less frequently) scan locking techniques as well.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**CMOS IMPLEMENTATION OF SWITCHING LATTICES**
Speaker: Levent Aksoy, Istanbul TU, TR
Authors: Ismail Cevik, Levent Aksoy and Mustafa Altun, Istanbul TU, TR

**Abstract**
Switching lattices consisting of four-terminal switches are introduced as area-efficient structures to realize logic functions. Many optimization algorithms have been proposed, including exact ones, realizing logic functions on lattices with the fewest number of four-terminal switches, as well as heuristic ones. Hence, the computing potential of switching lattices has been justified adequately in the literature. However, the same thing cannot be said for their physical implementation. There have been conceptual ideas for the technology development of switching lattices, but no concrete and directly applicable technology has been proposed yet. In this study, we show that switching lattices can be directly and efficiently implemented using a standard CMOS process. To realize a given logic function on a switching lattice, we propose static and dynamic logic solutions. The proposed circuits as well as the compared conventional ones are designed and simulated in the Cadence environment using TSMC 65nm CMOS process. Experimental post-layout results on logic functions show that switching lattices occupy much smaller area than those of the conventional CMOS implementations, while they have competitive delay and power consumption values.

**Download Paper (PDF; Only available from the DATE venue WiFi)**
IP1-3 A TIMING UNCERTAINTY-AWARE CLOCK TREE TOPOLOGY GENERATION ALGORITHM FOR SINGLE FLUX QUANTUM CIRCUITS
Speaker: Massoud Pedram, University of Southern California, US
Authors: Soheil Nazar Shahsavani, Bo Zhang and Massoud Pedram, University of Southern California, US
Abstract
This paper presents a low-cost, timing uncertainty-aware synchronous clock tree topology generation algorithm for single flux quantum (SFQ) logic circuits. The proposed method considers the criticality of the data paths in terms of timing slack as well as the total wirelength of the clock tree and generates a (height-1)-balanced binary clock tree using a bottom-up approach and an integer linear programming (ILP) formulation. The statistical timing analysis results for ten benchmark circuits show that the proposed method improves the total wirelength and the total negative hold slack by 4.2% and 64.6%, respectively, on average, compared with a wirelength-driven state-of-the-art balanced topology generation approach.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP1-4 SYMMETRY-BASED A/H-S BIST (SYM-BIST): DEMONSTRATION ON A SAR ADC IP
Speaker: Antonios Pavlidis, Sorbonne Université, CNRS, LIP6, FR
Authors: Antonios Pavlidis1, Marie-Minerve Louerat1, Eric Faehn2, Anand Kumar3 and Haralampos-G. Stratigopoulos1
1Sorbonne Université, CNRS, LIP6, FR; 2STMicroelectronics, FR; 3STMicroelectronics, IN
Abstract
In this paper, we propose a defect-oriented Built-In Self-Test (BIST) paradigm for analog and mixed-signal (A/MS) Integrated Circuits (ICs), called symmetry-based BIST (Sym-BIST). Sym-BIST exploits inherent symmetries in the design to generate invariances that should hold true only in defect-free operation. Violation of any of these invariances points to defect detection. We demonstrate SymBIST on a 65nm 10-bit Successive Approximation Register (SAR) Analog-to-Digital Converter (ADC) IP by ST Microelectronics. SymBIST does not result in any performance penalty, it incurs an area overhead of less than 5%, the test time equals about 16x the time to convert an analog input sample, it can be interfaced with a 2-pin digital access mechanism, and it covers the entire A/H-S part of the IP achieving a likelihood-weighted defect coverage higher than 85%.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP1-5 RANGE CONTROLLED FLOATING-GATE TRANSISTORS: A UNIFIED SOLUTION FOR UNLOCKING AND CALIBRATING ANALOG ICS
Speaker: Yiorgos Makris, University of Texas at Dallas, US
Authors: Sai Govinda Rao Nimmalapudi, Georgios Volanis, Yichuan Lu, Angelos Antonopoulos, Andrew Marshall and Yiorgos Makris, University of Texas at Dallas, US
Abstract
Analog Floating-Gate Transistors (AGFTs) are commonly used to fine-tune the performance of analog integrated circuits (ICs) after fabrication, thereby enabling high yield despite component mismatch and variability in semiconductor manufacturing. In this work, we propose a methodology that leverages such AGFTs to also prevent unauthorized use of analog ICs. Specifically, we introduce a locking mechanism that limits programming of AGFTs to a range which is inadequate for achieving the desired analog performance. Accordingly, our solution entails a two-step unlock-&-calibrate process. In the first step, AGFTs must be programmed through a secret sequence of voltages within that range, called waypoints. Successful following the waypoints unlocks the ability to program the AGFTs over their entire range. Thereby, in the second step, the typical AFGT-based post-silicon calibration process can be applied to adjust the performance of the IC within its specifications. Protection against brute-force or intelligent attacks attempting to guess the unlocking sequence is ensured through the vast space of possible waypoints in the continuous (analog) domain. Feasibility and effectiveness of the proposed solution is demonstrated and evaluated on an Operational Transconductance Amplifier (OTA). To our knowledge, this is the first solution which leverages the power of analog keys and addresses both unlocking and calibration needs of analog ICs in a unified manner.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP1-6 TESTING THROUGH SILICON VIAS IN POWER DISTRIBUTION NETWORK OF 3D-IC WITH MANUFACTURING VARIABILITY CANCELLATION
Speaker: Koutaro Hachiya, Teikyo Heisei University, JP
Authors: Koutaro Hachiya1 and Atsushi Kurokawa2
1Teikyo Heisei University, JP; 2Hiroasaki University, JP
Abstract
To detect open defects of power TSVs (Through Silicon Vias) in PDNs (Power Distribution Networks) of stacked 3D-ICs, a method was proposed which measures resistances between power macro-bumps to PDN and detects defects of TSVs by changes of the resistances. It suffers from manufacturing variabilities and may fail if one micro-bump directly under each TSV (direct-type placement style) to maximize its diagnostic performance, but the performance was not enough for practical applications. A variability cancellation method was also devised to improve the diagnostic performance. In this paper, a novel middle-type placement style is proposed which places one micro-bump between each pair of TSVs. Experimental simulations using a 3D-IC example show that the diagnostic performances of both the direct-type and the middle-type examples are improved by the variability cancellation and reach the practical level. The middle-type example outperforms the direct-type example in terms of number of micro-bumps and number of measurements.
Download Paper (PDF; Only available from the DATE venue WiFi)

IP1-7 TFAPPROX: TOWARDS A FAST EMULATION OF DNN APPROXIMATE HARDWARE ACCELERATORS ON GPU
Speaker: Zdenek Vasichek, Brno University of Technology, CZ
Authors: Filip Veverka, Vojtech Mrzcek, Zdenek Vasichek and Lukas Sekanina, Brno University of Technology, CZ
Abstract
Energy efficiency of hardware accelerators of deep neural networks (DNN) can be improved by introducing approximate arithmetic circuits. In order to quantify the error introduced by using these circuits and avoid the expensive hardware prototyping, a software emulator of the DNN accelerator is usually executed on CPU or GPU. However, this emulation is typically two or three orders of magnitude slower than a software DNN implementation running on CPU or GPU and operating with standard floating point arithmetic instructions and common DNN libraries. The reason is that there is no hardware support for approximate arithmetic operations on common CPUs and GPUs and these operations have to be expensive emulated. In order to address this issue, we propose an efficient emulation method for approximate circuits utilized in a given DNN accelerator which is emulated on GPU. All relevant approximate circuits are implemented as look-up tables and accessed through a texture memory mechanism of CUDA capable GPUs. We exploit the fact that the texture memory is optimized for irregular read-only access and in some GPU architectures is even implemented as a dedicated cache. This technique allowed us to reduce the inference time of the emulated DNN accelerator approximately 200 times with respect to an optimized CPU version on complex DNNs such as ResNet. The proposed approach extends the TensorFlow library and is available online at https://github.com/ehw-fit/tf-approximate
Download Paper (PDF; Only available from the DATE venue WiFi)
**[P1-8] BINARY LINEAR ECCS OPTIMIZED FOR BIT INVERSION IN MEMORIES WITH ASYMMETRIC ERROR PROBABILITIES**

**Speaker:** Valentin Gherman, CEA, FR  
**Authors:** Valentin Gherman, Samuel Evain and Bastien Giraud, CEA, FR  
**Abstract**  
Many memory types are asymmetric with respect to the error vulnerability of stored 0’s and 1’s. For instance, DRAM, STT-MRAM and NAND flash memories may suffer from asymmetric error rates. A recently proposed error-protection scheme consists in the inversion of the memory words with too many vulnerable values before they are stored in an asymmetric memory. In this paper, a method is proposed for the optimization of systematic binary linear block error-correcting codes in order to maximize their impact when combined with memory word inversion.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**[P1-9] BELDPC: BIT ERRORS AWARE ADAPTIVE RATE LDPC CODES FOR 3D TLC NAND FLASH MEMORY**

**Speaker:** Meng Zhang, Huazhong University of Science & Technology, CN  
**Authors:** Meng Zhang, Fei Wu, Qin Yu, Weihua Liu, Lanlan Cui, Yahui Zhao and Changsheng Xie, Huazhong University of Science & Technology, CN  
**Abstract**  
Three-dimensional (3D) NAND flash memory has high capacity and cell storage density by using the multi-bit technology and vertical stack architecture, but degrading data reliability due to high raw bit error rates (RBER) caused by program/erase (P/E) cycles and retention periods. Low-density parity-check (LDPC) codes become more popular error-correcting technologies to improve data reliability due to strong error correction capability, but introducing more decoding iterations at higher RBER. To reduce decoding iterations, this paper proposes BelDPC: bit errors aware adaptive rate LDPC codes for 3D triple-level cell (TLC) NAND flash memory. Firstly, bit error characteristics in 3D charge trap TLC NAND flash memory are studied on a real FPGA testing platform, including asymmetric bit flipping and temporal locality of bit errors. Then, based on these characteristics, a high-efficiency LDPC code is designed. Experimental results show BelDPC can reduce decoding iterations under different P/E cycles and retention periods.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**[P1-10] POISONING THE (DATA) WELL IN ML-BASED CAD: A CASE STUDY OF HIDING LITHOGRAPHIC HOTSPOTS**

**Speaker:** Kang Liu, New York University, US  
**Authors:** Kang Liu, Benjamin Tan, Ramesh Karri and Siddharth Garg, New York University, US  
**Abstract**  
Machine learning (ML) provides state-of-the-art performance in many parts of computer-aided design (CAD) flows. However, deep neural networks (DNNs) are susceptible to various adversarial attacks, including data poisoning to compromise training to insert backdoors. Sensitivity to training data integrity presents a security vulnerability, especially in light of malicious insiders who want to cause targeted neural network misbehavior. In this study, we explore this threat in lithographic hotspot detection via training data poisoning, where hotspots in a layout clip can be “hidden” at inference time by including a trigger shape in the input. We show that training data poisoning attacks are feasible and stealthy, demonstrating a backdoored neural network that performs normally on clean inputs but misbehaves on inputs when a backdoor trigger is present. Furthermore, our results raise some fundamental questions about the robustness of ML-based systems in CAD.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**[P1-11] SOLOMON: AN AUTOMATED FRAMEWORK FOR DETECTING FAULT ATTACK VULNERABILITIES IN HARDWARE**

**Speaker:** Milind Srivastava, IIT Madras, IN  
**Authors:** Milind Srivastava, Patanjali SLPShk, Indrani Roy, Chester Rebeiro, Aritra Hazra and Swarup Bhunia  
IIT Madras, IN; IIT Kharagpur, IN; University of Florida, US  
**Abstract**  
Fault attacks are potent physical attacks on crypto-devices. A single fault injected during encryption can reveal the cipher’s secret key. In a hardware realization of an encryption algorithm, only a tiny fraction of the gates is exploitable by such an attack. Finding these vulnerable gates has been a manual and tedious task requiring considerable expertise. In this paper, we propose SOLOMON, the first automatic fault attack vulnerability detection framework for hardware designs. Given a cipher implementation, either at RTL or gate-level, SOLOMON uses formal methods to map vulnerable regions in the cipher algorithm to specific locations in the hardware thus enabling targeted countermeasures to be deployed with much lesser overheads. We demonstrate the efficacy of the SOLOMON framework using three ciphers: AES, CLEFIA, and Simon.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**[P1-12] FORMAL SYNTHESIS OF MONITORING AND DETECTION SYSTEMS FOR SECURE CPS IMPLEMENTATIONS**

**Speaker:** Ipsita Koley, IIT Kharagpur, IN  
**Authors:** Ipsita Koley, Saurav Kumar Ghosh, Dey Soumyajit, Debdeep Mukhopadhyay, Amogh Kashyap K N, Sachin Kumar Singh, Lavenya Lokesh, Jithin Nalu Purakkal and Nishant Sinha  
IIT Kharagpur, IN; Robert Bosch Engineering and Business Solutions Private Limited, IN  
**Abstract**  
We consider the problem of securing a given control loop implementation of a cyber-physical system (CPS) in the presence of Man-in-the-Middle attacks on data exchange between plant and controller over a compromised network. To this end, there exists various detection schemes which provide mathematical guarantees against such attacks for the theoretical control model. However, such guarantees may not hold for the actual control software implementation. In this article, we propose a formal approach towards synthesizing attack detectors with varying thresholds which can prevent performance degrading stealthy attacks while minimizing false alarms.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**[P1-13] ASCELLA: ACCELERATING SPARSE COMPUTATION BY ENABLING STREAM ACCESESSES TO MEMORY**

**Speaker:** Bahar Asgari, Georgia Tech, US  
**Authors:** Bahar Asgari, Ramyad Hedidi and Hyeson Kim, Georgia Tech, US  
**Abstract**  
Sparse computations dominate a wide range of applications from scientific problems to graph analytics. The main characterization of sparse computations, indirect memory accesses, prevents them from effectively achieving high performance on general-purpose processors. Therefore, hardware accelerators have been proposed for sparse problems. For these accelerators, the storage format and the decompression mechanism is crucial but have been less attention in prior work. To address this gap, we propose Ascella, an accelerator for sparse computations, which enables a seamless stream of data and parallel computation, proposes a fast decompression mechanism. Our implementation on a ZYNQ FPGA shows that on average, Ascella executes sparse problems up to 5.1x as fast as prior work.

**Download Paper (PDF; Only available from the DATE venue WiFi)**
ACCELERATION OF PROBABILISTIC REASONING THROUGH CUSTOM PROCESSOR ARCHITECTURE
Speaker: Nimish Shah, KU Leuven, BE
Authors: Nimish Shah, Laura I. Galindez Olascoaga, Wannes Meert and Marian Verhelst, KU Leuven, BE
Abstract: Probabilistic reasoning is an essential tool for robust decision-making systems because of its ability to explicitly handle real-world uncertainty, constraints and causal relations. Consequently, researchers are developing hybrid models by combining Deep Learning with Probabilistic reasoning for safety-critical applications like self-driving vehicles, autonomous drones, etc. However, probabilistic reasoning kernels do not execute efficiently on CPUs or GPUs. This paper, therefore, proposes a custom programmable processor to accelerate sum-product networks, an important probabilistic reasoning execution kernel. The processor has an optimized datapath architecture and memory hierarchy optimized for sum-product networks execution. Experimental results show that the processor, while requiring fewer computational and memory units, achieves a 12x throughput benefit over the Nvidia Jetson TX2 embedded GPU platform.
Download Paper (PDF; Only available from the DATE venue WiFi)

A PERFORMANCE ANALYSIS FRAMEWORK FOR REAL-TIME SYSTEMS SHARING MULTIPLE RESOURCES
Speaker: Shayan Tabatabaei Nikkhah, Eindhoven University of Technology, NL
Authors: Shayan Tabatabaei Nikkhah1, Marc Geilen2, Dip Goswami1 and Kees Goossens2
1Eindhoven University of Technology, NL; 2Eindhoven university of technology, NL
Abstract: Timing properties of applications strongly depend on resources that are allocated to them. Applications often have multiple resource requirements, all of which must be met for them to proceed. Performance analysis of event-based systems has been widely studied in the literature. However, the proposed works consider only one resource requirement for each application task. Additionally, they mainly focus on the rate at which resources serve applications (e.g., power, instructions or bits per second), but another aspect of resources, which is their provided capacity (e.g., energy, memory ranges, FPGA regions), has been ignored. In this work, we propose a mathematical framework to describe the provisioning rate and capacity of various types of resource. Additionally, we consider the simultaneous use of multiple resources. Conservative bounds on response times of events and their backlog are computed. We prove that the bounds are monotone in event arrivals and in required and provided rate and capacity, which enables verification of real-time application performance based on worst-case characterizations. The applicability of our framework is shown in a case study.
Download Paper (PDF; Only available from the DATE venue WiFi)

SCALING UP THE MEMORY INTERFERENCE ANALYSIS FOR HARD REAL-TIME MANY-CORE SYSTEMS
Speaker: Matheus Schuh, Verimag / Kalray, FR
Authors: Matheus Schuh1, Maximilien Dupont de Dinechin2, Matthieu Moy3 and Claire Maiza4
1Verimag / Kalray, FR; 2ENS Paris / ENS Lyon / LIP, FR; 3ENS Lyon / LIP, FR; 4Grenoble INP / Verimag, FR
Abstract: In RTNS 2016, Rihani et al. proposed an algorithm to compute the impact of interference on memory accesses on the timing of a task graph. It calculates a static, time-triggered schedule, i.e. a release date and a worst-case response time for each task. The task graph is a DAG, typically obtained by compilation of a high-level dataflow language, and the tool assumes a previously determined mapping and execution order. The algorithm is precise, but suffers from a high O(n^4) complexity, n being the number of input tasks. Since we target many-core platforms with tens or hundreds of cores, applications likely to exploit the parallelism of these platforms are too large to be handled by this algorithm in reasonable time. This paper proposes a new algorithm that solves the same problem. Instead of performing global fixed-point iterations on the task graph, we compute the static schedule incrementally, reducing the complexity to O(n^2). Experimental results show a reduction from 535 seconds to 0.90 seconds on a benchmark with 384 tasks, i.e. 593 times faster.
Download Paper (PDF; Only available from the DATE venue WiFi)

A PERFORMANCE ANALYSIS FRAMEWORK FOR REAL-TIME SYSTEMS SHARING MULTIPLE RESOURCES
Speaker: Prabhat Mishra, University of Florida, US
Authors: Subodha Charles, Megan Logan and Prabhat Mishra, University of Florida, US
Abstract: System-on-Chip (SoC) supply chain is widely acknowledged as a major source of security vulnerabilities. Potentially malicious third-party IPs integrated on the same Network-on-Chip (NoC) with the trusted components can lead to security and trust concerns. While secure communication is a well studied problem in computer networks domain, it is not feasible to implement those solutions on resource-constrained SoCs. In this paper, we present a lightweight anonymous routing protocol for communication between IP cores in NoC based SoCs. Our method eliminates the major overhead associated with traditional anonymous routing protocols while ensuring that the desired security goals are met. Experimental results demonstrate that existing security solutions on NoC can introduce significant (1.5X) performance degradation, whereas our approach provides the same security features with minor (4%) impact on performance.
Download Paper (PDF; Only available from the DATE venue WiFi)

A NON-INVASIVE WEARABLE BIDIMPEDEANCE SYSTEM TO WIRELESSLY MONITOR BLADDER FILLING
Speaker: Michele Magno, ETH Zurich, CH
Authors: Markus Reichmuth, Simone Schuerle and Michele Magno, ETH Zurich, CH
Abstract: Monitoring of renal function can be crucial for patients in acute care settings. Commonly during postsurgical surveillance, urinary catheters are employed to assess the urine output accurately. However, as with any external device inserted into the body, the use of these catheters carries a significant risk of infection. In this paper, we present a non-invasive method to measure the fill rate of the bladder, and thus the rate of renal clearance, via an external bioimpedance sensor system to avoid the use of urinary catheters, thereby eliminating the risk of infections and improving patient comfort. We design and propose a 4-electrode front-end and the whole wearable and wireless system with low power and accuracy in mind. The results demonstrate the accuracy of the sensors and low power consumption of only 80µW with a duty cycling of 1 acquisition every 5 minutes, which makes this battery-operated wearable device a long-term monitor system.
Download Paper (PDF; Only available from the DATE venue WiFi)
### 4.1 Hardware-enabled security

**Date:** Tuesday 10 March 2020  
**Time:** 17:00 - 18:30  
**Location / Room:** Amphithéâtre Jean Prouve  
**Chair:** Marchand Cedric, Ecole Centrale de Lyon, FR  
**Co-Chair:** Hai Zhou, Northwestern University, US

This session covers solutions in hardware-based design to improve security. The papers in the session propose a NTT (Number Theoretic Transform) technique enabling faster polynomial multiplication, a reliable key-PUF for key generation, and a runtime circuit de-obfuscating solution. Post-Quantum cryptography and new attacks will be discussed along this session.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>17:00</td>
<td>4.1.1</td>
<td><strong>A FLEXIBLE AND SCALABLE NTT HARDWARE: APPLICATIONS FROM HOMOMORPHICALLY ENCRYPTED DEEP LEARNING TO POST-QUANTUM CRYPTOGRAPHY</strong></td>
<td>Ahmet Can Mert, Sabanci University, TR; Ahmet Can Mert, Erdinc Ozturk, Erkay Savas, Michela Becchi, Aydin Aysu, Sabanci University, TR; North Carolina State University, US</td>
</tr>
</tbody>
</table>

**Abstract**  
The Number Theoretic Transform (NTT) enables faster polynomial multiplication and is becoming a fundamental component of next-generation cryptographic systems. NTT hardware designs have two prevalent problems related to design-time flexibility. First, algorithms have different arithmetic structures causing the hardware designs to be manually tuned for each setting. Second, applications have diverse throughput/area needs but the hardware have been designed for a fixed, pre-defined number of processing elements. This paper proposes a parametric NTT hardware generator that takes arithmetic configurations and the number of processing elements as inputs to produce an efficient hardware with the desired parameters and throughput. We illustrate the employment of the proposed design in two applications with different needs: A homomorphically encrypted deep neural network inference (CryptoNets) and a post-quantum digital signature scheme (qTESLA). We propose the first NTT hardware accelerator for both applications on FPGAs. Compared to prior software and high-level synthesis solutions, the results show that our hardware can accelerate NTT up to 28x and 48x, respectively. Therefore, our work paves the way for high-level, automated, and modular design of next-generation cryptographic hardware solutions.

Download Paper (PDF; Only available from the DATE venue WiFi)

| 17:30  | 4.1.2 | **RELIABLE AND LIGHTWEIGHT PUF-BASED KEY GENERATION USING VARIOUS INDEX VOTING ARCHITECTURE** | Jeong-Hyeon Kim, Sungkyunkwan University, KR; Jeong-Hyeon Kim, Ho-Jun Jo, Kyung-kuk Jo, Sunghee Cho, Jaeyong Chung, and Joon-Sung Yang, Sungkyunkwan University, KR; Incheon National University, KR |

**Abstract**  
Physical Unclonable Functions (PUFs) can be utilized for secret key generation in security applications. Since the inherent randomness of PUF can degrade its reliability, most of the existing PUF architectures have designed post-processing logic to enhance the reliability such as an error correction function for guaranteeing reliability. However, the structures incur high cost in terms of implementation area and power consumption. This paper introduces a Various Index Voting Architecture (VIVA) that can enhance the reliability with a low overhead compared to the conventional schemes. The proposed architecture is based on an index-based scheme with simple computation logic units and iterative operations to generate multiple indices for the accuracy of key generation. Our evaluation results show that the proposed architecture reduces the hardware implementation overhead by 2 to more than 5 times, without losing a key generation failure probability compared to conventional approaches.

Download Paper (PDF; Only available from the DATE venue WiFi)
ESTIMATING THE CIRCUIT DE-OBFUSCATION RUNTIME BASED ON GRAPH DEEP LEARNING

Speaker: Gaurav Kolhe, George Mason University, US

Authors: Zhiqian Chen1, Gaurav Kolhe, Setareh Rafatrad2, Chang-Tien Lu2, Sai Manoj Pudukotai Dinakarrao2, Houman Homayoun2 and Liang Zhao2
1Virginia Tech, US; 2George Mason University, US

Abstract
Random values from discrete distributions are typically generated from uniformly-random samples. A common technique is to use a cumulative distribution table (CDT) lookup for inversion sampling, but it is also possible to use Boolean functions to map a uniformly-random bit sequence into a value from a discrete distribution. This work presents a methodology for deriving such functions for any discrete distribution, encoding them in VHDL for implementation in combinational hardware, and (for moderate precision and sample space size) confirming the correctness of the produced distribution. The process is demonstrated using a discrete Gaussian distribution with a small sample space, but it is applicable to any discrete distribution with fixed parameters.

Results
The process produces compact solutions for distributions up to moderate size and precision.

Download Paper (PDF; Only available from the DATE venue WiFi)

ON THE PERFORMANCE OF NON-PROFILED DIFFERENTIAL DEEP LEARNING ATTACKS AGAINST AN AES CRYPTOGRAPHY

Speaker: Amir Alipour, Grenoble INP Esisar, FR

Authors: Amir Alipour1, Athanasios Papadimitriou2, Vincent Berouillé3, Ehsan Aerabi3 and David Hel3
1University Grenoble Alpes, Grenoble INP ESISAR, LCIS Laboratory, FR; 2University Grenoble Alpes, Grenoble INP ESISAR, ESYNOV, FR; 3University Grenoble Alpes, Grenoble INP ESISAR, LSIC Laboratory, FR

Abstract
Recent works in the field of cryptography focus on Deep Learning based Side Channel Analysis (DLSCA) as one of the most powerful attacks against common encryption algorithms such as AES. As a common case, profiling DLSCA have shown great capabilities in revealing secret cryptographic keys against the majority of AES implementations. In a very recent study, it has been shown that Deep Learning can be applied in a non-profiling way (non-profiling DLSCA), making this method considerably more practical, and able to break powerful countermeasures for encryption algorithms such as AES including masking countermeasures, requiring considerably less power traces than a first order CPA attack. In this work, our main goal is to apply the non-profiling DLSCA against a hiding-based AES countermeasure which utilizes correlated noise generation so as to hide the secret encryption key. We show that this AES, with correlated noise generation as a lightweight countermeasure, can provide equivalent protection under CPA and under non-profiling DLSCA attacks, in terms of the required power traces to obtain the secret key.

Download Paper (PDF; Only available from the DATE venue WiFi)
<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
</table>
| 17:00 | 4.2.1 | FAST AND ACCURATE DRAM SIMULATION: CAN WE FURTHER ACCELERATE IT? | Matthias Jung, Fraunhofer IESE, DE  
Johannes Feldmann, Matthias Jung, Kira Kraft, Lukas Steiner and Norbert Wehn  
1TU Kaiserslautern, DE; 2Fraunhofer IESE, DE |
| 17:30 | 4.2.2 | ACCURATE AND EFFICIENT CONTINUOUS TIME AND DISCRETE EVENTS SIMULATION IN SYSTEMC | Breitner Fernandez-Mesa, TIMA Laboratory, University Grenoble Alpes, FR  
Authors:  
Breitner Fernandez-Mesa, Liliana Andrade and Frédéric Pétrot, TIMA Lab, Université Grenoble Alpes, FR |
| 18:00 | 4.2.3 | MODELING AND VERIFYING UNCERTAINTY-AWARE TIMING BEHAVIORS USING PARAMETRIC LOGICAL TIME CONSTRAINT | Fei Gao, East China Normal University, CN  
Authors:  
Fei Gao, Mallet Frederic, Min Zhang and Mingsong Chen  
1East China Normal University, CN; 2Université Cote d’Azur, CNRS, Inria, I3S, Nice, France, FR; 3East China Normal University, FR |
| 18:30 | IP2-3, 831 | FAST AND ACCURATE PERFORMANCE EVALUATION FOR RISC-V USING VIRTUAL PROTOTYPES | Vladimir Herdt, University of Bremen, DE  
Authors:  
Vladimir Herdt, Daniel Grossel and Rolf Drechsler  
1University of Bremen, DE; 2University of Bremen / DFKI, DE |
| 18:31 | IP2-4, 641 | AUTOMATED GENERATION OF LTL SPECIFICATIONS FOR SMART HOME IOT USING NATURAL LANGUAGE | Shiyu Zhang, Nanjing University, CN  
Authors:  
Shiyu Zhang, Juan Zhai, Lei Bu, Mingsong Chen, Linzhang Wang and Xuandong Li  
1Nanjing University, CN; 2East China Normal University, CN |

### 4.3 EU Projects on Nanoelectronics with CMOS and alternative technologies

**Date:** Tuesday 10 March 2020  
**Time:** 17:00 – 18:30  
**Location / Room:** Autrans  
**Chair:** Dimitris Gizopoulos, UoA, GR
This session presents the results of three European Projects in different stages of execution covering the development of a complete synthesis and optimization methodology for nano-crossbar arrays; the reliability, security, and associated EDA tools for nanoelectronic systems, and the exploitation of STT-MTJ technologies for heterogeneous function implementation.

**Time** | **Label** | **Presentation Title** | **Authors**
--- | --- | --- | ---
18:00 | 4.3.1 | **NANO-CROSSBAR BASED COMPUTING: LESSONS LEARNED AND FUTURE DIRECTIONS** | MustaA Altun, Istanbul TU, TR
**Abstract**
In this paper, we first summarize our research activities done through our European Union's Horizon-2020 project between 2015 and 2019. The project has a goal of developing synthesis and performance optimization techniques for nanocrossbar arrays. For this purpose, different computing models including diode, memristor, FET, and four-terminal switch based models, within different technologies including carbon nanotubes, nanowires, and memristors as well as the CMOS technology have been investigated. Their capabilities to realize logic functions and to tolerate faults have been deeply analyzed. From these experiences, we think that instead of replacing CMOS with a completely new crossbar based technology, developing CMOS compatible crossbar technologies and computing models is a more viable solution to overcome challenges in CMOS miniaturization. At this point, four terminal switch based arrays, called switching lattices, come forward with their CMOS compatibility feature as well as with their area efficient device and circuit realizations. We have showed that switching lattices can be efficiently implemented using a standard CMOS process to implement logic functions by doing experiments in a 65nm CMOS process. Further in this paper, we make an introduction of realizing memory arrays with switching lattices including ROMs and RAMs. Also we discuss challenges and promises in realizing switching lattices for under 30nm CMOS technologies including FinFET technologies.
--- | 4.3.2 | **RESCUE: INTERDEPENDENT CHALLENGES OF RELIABILITY, SECURITY AND QUALITY IN NANOELECTRONIC SYSTEMS** | Maksim Jeninhi, Tallinn University of Technology, EE
**Abstract**
The recent trends for nanoelectronic computing systems include machine-to-machine communication in the era of Internet-of-Things (IoT) and autonomous systems, complex safety-critical applications, extreme miniaturization of implementation technologies and intensive interaction with the physical world. These set tough requirements on mutually dependent extra-functional design aspects. The H2020 MSCA ITN project RESCUE is focused on key challenges for reliability, security and quality, as well as related electronic design automation tools and methodologies. The objectives include both research advancements and cross-sectoral training of a new generation of interdisciplinary researchers. Notable interdisciplinary collaborative research results for the first half-period include novel approaches for test generation, soft-error and transient faults vulnerability analysis, cross-layer fault-tolerance and error-resilience, functional safety validation, reliability assessment and run-time management, HW security enhancement and initial implementation of these into holistic EDA tools.
--- | 4.3.3 | **A UNIVERSAL SPINTRONIC TECHNOLOGY BASED ON MULTIFUNCTIONAL STANDARDIZED STACK** | Mehdi Tahoori, Karlsruhe Institute of Technology, DE
**Abstract**
The goal of the GREAT RIA project is to co-integrate multiple functions like sensors ("Sensing"), RF emitters or receivers ("Communicating") and logic/memory ("Processing/Storing") together within CMOS by adapting the Spin-Transfer Torque Magnetic Tunnel Junction (STT-MTJ), elementary constitutive cell of the MRAM memories, to a single baseline technology. Based on the STT unique set of performances (non-volatility, high speed, infinite endurance and moderate read/write power), GREAT will achieve the same goal as heterogeneous integration of devices but in a much simpler way. This will lead to a unique STT-MTJ cell technology called Multifunctional Standardized Stack (MSS). This paper presents the lessons learned in the project from the technology, compact modeling, process design kit, standard cells, as well as memory and system level design evaluation and exploration. The proposed technology and toolsets are giant leaps towards heterogeneous integrated technology and architectures for IoT.

**Time** | **Label** | **Presentation Title** | **Authors**
--- | --- | --- | ---
18:30 | End of session | | |
A LEARNING-BASED THERMAL SIMULATION FRAMEWORK FOR EMERGING TWO-PHASE COOLING TECHNOLOGIES

Speaker: Ayse Coskun, Boston University, US

Authors: Ayse Coskun1, Geoffrey Vaartstra2, Prachi Shukla1, Zhengmao Lu2, Evelyn Wang2, Sherief Reda1 and Ayse Coskun1

Abstract: Future high-performance chips will require new cooling technologies that can extract heat efficiently. Two-phase cooling is a promising processor cooling solution owing to its high heat transfer rate and potential benefits in cooling power. Two-phase cooling mechanisms, including microchannel-based two-phase cooling or two-phase vapor chambers (VCs), are typically modeled by computing the temperature-dependent heat transfer coefficient (HTC) of the evaporator or coolant using an iterative simulation framework. Precomputed HTC correlations are specific to a given cooling system design and cannot be applied to even the same cooling technology with different cooling parameters (such as different geometries). Another challenge is that HTC correlations are typically calculated with computational fluid dynamics (CFD) tools, which induce long design and simulation times. This paper introduces a learning-based temperature-dependent HTC simulation framework that is used to model a two-phase cooling solution with a wide range of cooling design parameters. In particular, the proposed framework includes a compact thermal model (CTM) of two-phase VCs with hybrid wick evaporators (of nanoporous membrane and microchannels). We build a new simulation tool to integrate the proposed simulation framework and CTM. We validate the proposed simulation framework as well as the new CTM through comparisons against a CFD model. Our simulation framework and CTM achieve a speedup of 21X with an average error of 0.98deg(C) (and a maximum error of 2.59degC). We design an optimization flow for hybrid wicks to select the most beneficial nanoporous membrane and microchannel geometries. Our flow is capable of finding a geometry-coolant combination that results in a lower (or similar) maximum chip temperature compared to that of the best coolant-geometry pair selected by grid search, while providing a speedup of 9.4X.

Download Paper (PDF; Only available from the DATE venue WiFi)

LIGHTWEIGHT THERMAL MONITORING IN OPTICAL NETWORKS-ON-CHIP VIA ROUTER REUSE

Speaker: Mengquan Li, Nanyang Technological University, US

Authors: Mengquan Li1, Jun Zhou2 and Weichen Liu2

Abstract: Optical network-on-chip (ONoC) is an emerging communication architecture for manycore systems due to low latency, high bandwidth, and low power dissipation. However, a major concern lies in its thermal susceptibility -- under on-chip temperature variations, functional nanophotonic devices, especially microring resonator (MR)-based devices, suffer from significant thermal-induced optical power loss, which may counteract the power advantages of ONoCs and even cause functional failures. Considering the fact that temperature gradients are typically found on many-core systems, effective thermal monitoring, performing as the foundation of thermal-aware management, is critical on ONoCs. In this paper, a lightweight thermal monitoring scheme is proposed for ONoCs. We first design a temperature measurement module based on generic optical routers. It introduces trivial overheads in chip area by reusing the components in routers. A major problem with reusing optical routers is that it may potentially interfere with the normal communications in ONoCs. To address it, we then propose a time allocation strategy to schedule thermal sensing operations in the time intervals between communications. Evaluation results show that our scheme exhibits an untrimmed inaccuracy of 1.0070 K with low energy consumption of 656.38 pJ/Sa. It occupies an extremely small area of 0.0020 mm^2, reducing the area cost by 83.74% on average compared to the state-of-the-art optical thermal sensor design.

Download Paper (PDF; Only available from the DATE venue WiFi)

A SPECTRAL APPROACH TO SCALABLE VECTORLESS THERMAL INTEGRITY VERIFICATION

Speaker: Zhuo Feng, Stevens Institute of Technology, US

Authors: Zhiqiang Zhao1, Mengquan Li2, and Zhuo Feng2

Abstract: Existing chip thermal analysis and verification methods require detailed distribution of power densities or modeling of underlying input workloads (vectors), which may not always be feasible at early-design stage. This paper introduces the first vectorless thermal integrity verification framework that allows computing worst-case temperature (gradient) distributions across the entire chip under a set of local and global workload (power density) constraints. To address the computational challenges introduced by the large 3D mesh-structured thermal grids, we propose a novel spectral approach for highly-scalable vectorless thermal verification of large chip designs. Our approach is based on emerging spectral graph theory and graph signal processing techniques, which consists of a thermal grid topology sparsification phase, an edge weight scaling phase, as well as a solution refinement procedure. The effectiveness and efficiency of our approach have been demonstrated through extensive experiments.

Download Paper (PDF; Only available from the DATE venue WiFi)

DYNAMIC THERMAL MANAGEMENT WITH PROACTIVE FAN SPEED CONTROL THROUGH REINFORCEMENT LEARNING

Speaker: Arman Iranfar, EPFL, CH

Authors: Arman Iranfar1, Federico Terraneo2, Gabor Csordas1, Marina Zapater1, William Fornaciari2 and David Atienza1

Abstract: Dynamic Thermal Management (DTM) in submicron technology has become a major challenge since it directly affects Multiprocessors Systems-on-chip (MPSoCs) performance, power consumption, and lifetime reliability. For proper DTM, thermal simulators play a significant role as they allow chip temperature to be safely studied. Nonetheless, state-of-the-art thermal simulators do not support transient fan models. As a result, adaptive fan speed control, which is an important runtime parameter, cannot be well utilized in DTM. Therefore, in this work, we first propose and integrate a transient fan model into a state-of-the-art thermal simulator, enabling adaptive fan speed control simulation for efficient DTM. We, then, validate our simulation framework through a thermal test chip achieving less than 2\% error in the worst case. With multiple fan speeds, however, the DTM design space grows significantly, which can ultimately make conventional solutions, such as grid search, infeasible, impractical, or insufficient due to the large runtime overhead. Therefore, we address this challenge through a reinforcement learning-based solution to proactively determine number of active cores, operating frequency, and fan speed. The proposed solution is able to reduce fan power by up to 40% compared to a DTM with constant fan speed with less than 1% performance degradation. Also, compared to a state-of-the-art DTM technique our solution improves the performance by up to 19% for the same fan power.

Download Paper (PDF; Only available from the DATE venue WiFi)
## 4.5 Adaptation and optimization for real-time systems

**Date:** Tuesday 10 March 2020  
**Time:** 17:00 - 18:30  
**Location / Room:** Bayard

**Chair:** Wanli Chang, University of York, GB  
**Co-Chair:** Emmanuel Grolleau, ENSMA, FR

This session presents novel techniques for systems requiring adaptations. The papers in this session are including monitoring techniques to increase reactivity, considering weakly-hard constraints, extending previous cache persistence analyses from one core to several cores, and modeling data chains while latency bounds are ensured.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
<th>Authors</th>
<th>Authors</th>
<th>Authors</th>
</tr>
</thead>
</table>
| 18:30  | IP2-5 | A HEAT-RECIRCULATION-AWARE VM PLACEMENT STRATEGY FOR DATA CENTERS                  | Hao Feng¹, Yuhui Deng² and Yi Zhou³  
¹Jinan University, CN; ²Chinese Academy of Sciences; ³Jinan University, CN; ⁴Columbus State University, US  
**Abstract**  
Data centers consisted of a great number of IT devices (e.g., servers, switches and etc.) which generates a massive amount of heat emission. Due to the special arrangement of racks in the data center, heat recirculation often occurs due to the uneven node. It can cause a sharp rise in temperature of the equipment coupled with local hot spots in data centers. Existing VM placement strategies can minimize energy consumption of data centers by optimizing resource allocation in terms of multiple physical resources (e.g., memory, bandwidth, CPU and etc.). However, existing strategies ignore the role of heat recirculation in the data center. To address this problem, in this study, we propose a heat-recirculation-aware VM placement strategy and design a Simulated Annealing Based Algorithm (SABA) to lower the energy consumption of data centers. Different from the existing SA algorithm, SABA optimize the distribution of the initial solution and the way of iteration. We quantitatively evaluate SABA's performance in terms of algorithm efficiency, the activated servers and the energy saving against with XINT-GA algorithm (Thermal-aware task scheduling Strategy). FCS (First-Come First-Served), and SA. Experimental results indicate that our heat-recirculation-aware VM placement strategy provides a powerful solution for improving energy efficiency of data centers.  
**Download Paper (PDF; Only available from the DATE venue WiFi)** |
| 18:31  | IP2-6 | ENERGY OPTIMIZATION IN NCFET-BASED PROCESSORS                                     | Sami Salamin¹, Martin Rapp¹, Hussam Amrouch¹, Andreas Gerstlauer² and Joerg Henkel¹  
¹Karlsruhe Institute of Technology, DE; ²University of Texas at Austin, US  
**Abstract**  
Energy consumption is a key optimization goal for all modern processors. Negative Capacitance Field-Effect Transistors (NCFETs) are a leading emerging technology that promises outstanding performance addition to better energy efficiency. The thickness of the additional ferroelectrical layer, frequency, and voltage are the key parameters in NCFET technology that impact the power and frequency of processors. However, their joint impact on energy optimization has not been investigated yet. In this work, we are the first to demonstrate that conventional (i.e., NCFET-unaware) dynamic voltage/frequency scaling (DVFS) techniques to minimize energy are sub-optimal when applied to NCFET-based processors. We further demonstrate that state-of-the-art NCFET-aware voltage scaling for power minimization is also sub-optimal when it comes to energy. This work provides the first NCFET-aware DVFS technique that optimizes the processor's energy through optimal runtime frequency/voltage selection. In NCFETs, energy-optimal frequency and voltage are dependent on the workload and technology parameters. Our NCFET-aware DVFS technique considers these effects to perform optimal voltage/frequency selection at runtime depending on workload characteristics. Results show up to 90 % energy savings compared to conventional DVFS techniques. Compared to state-of-the-art NCFET-aware power management, our technique provides up to 72 % energy savings along with 3:7x higher performance.  
**Download Paper (PDF; Only available from the DATE venue WiFi)** |
| 18:30  | IP2-6 | ENERGY OPTIMIZATION IN NCFET-BASED PROCESSORS                                     | Sami Salamin¹, Martin Rapp¹, Hussam Amrouch¹, Andreas Gerstlauer² and Joerg Henkel¹  
¹Karlsruhe Institute of Technology, DE; ²University of Texas at Austin, US  
**Abstract**  
Energy consumption is a key optimization goal for all modern processors. Negative Capacitance Field-Effect Transistors (NCFETs) are a leading emerging technology that promises outstanding performance addition to better energy efficiency. The thickness of the additional ferroelectrical layer, frequency, and voltage are the key parameters in NCFET technology that impact the power and frequency of processors. However, their joint impact on energy optimization has not been investigated yet. In this work, we are the first to demonstrate that conventional (i.e., NCFET-unaware) dynamic voltage/frequency scaling (DVFS) techniques to minimize energy are sub-optimal when applied to NCFET-based processors. We further demonstrate that state-of-the-art NCFET-aware voltage scaling for power minimization is also sub-optimal when it comes to energy. This work provides the first NCFET-aware DVFS technique that optimizes the processor's energy through optimal runtime frequency/voltage selection. In NCFETs, energy-optimal frequency and voltage are dependent on the workload and technology parameters. Our NCFET-aware DVFS technique considers these effects to perform optimal voltage/frequency selection at runtime depending on workload characteristics. Results show up to 90 % energy savings compared to conventional DVFS techniques. Compared to state-of-the-art NCFET-aware power management, our technique provides up to 72 % energy savings along with 3:7x higher performance.  
**Download Paper (PDF; Only available from the DATE venue WiFi)** |
A PARTICLE SWARM OPTIMIZATION GUIDED APPROXIMATE KEY SEARCH ATTACK ON LOGIC LOCKING IN THE ABSENCE OF SCAN ACCESS

Speaker: Rajit Karmakar, IIT KHARAGPUR, IN
Authors: RAJIT KARMAKAR and Santanu Chattopadhayay, IIT Kharagpur, IN

Abstract

Logic locking is a well known Design-for-Security(DfS) technique for Intellectual Property (IP) protection of digital Integrated Circuits(IC). However, various attacks on logic locking can extract the secret obfuscation key successfully. Although Boolean Satisfiability (SAT) attacks can break most of the logic locked circuits, inability to deobfuscate sequential circuits is the main limitation of this type of attacks. Several existing defense strategies exploit this fact to thwart SAT attack by obfuscating the scan-based Design-for-Testability (DfT) infrastructure. In the absence of scan access, Model Checking based circuit unrolling attacks also suffer from scalability issues. In this paper, we propose a particle swarm optimization (PSO) guided attack framework, which is capable of finding an approximate key that produces correct output in most of the cases. Unlike the SAT attacks, the proposed attack framework can work even in the absence of scan access. Unlike Model Checking attacks, it does not suffer from scalability issues, thus can be applied on significantly large sequential circuits.

Experimental results show that the derived key can produce correct outputs in more than 99% cases, for the majority of the benchmark circuits, while for the rest of the circuits, a minimal error is observed. The proposed attack framework enables partial activation of large sequential circuits in the absence of scan access, which is not feasible using the existing attack frameworks.

Download Paper (PDF; Only available from the DATE venue WiFi)
EFFECT OF AGING ON PUF MODELING ATTACKS BASED ON POWER SIDE-CHANNEL OBSERVATIONS

Speaker: Trevor Kroeger, University of Maryland Baltimore County, US

Authors: Trevor Kroeger1, Wei Cheng2, Jean Luc Danger3, Sylvain Guilley4 and Naghmeh Karimi5

1University of Maryland Baltimore County, US; 2Telecom ParisTech, FR; 3Telecom ParisTech, FR; 4Secure-IC, FR; 5University of Maryland, Baltimore County, US

Abstract

Thanks to the imperfections in manufacturing process, Physically Unclonable Functions (PUFs) produce their unique outputs for given input signals (challenges) fed to identical circuitry designs. PUFs are often used as hardware primitives to provide security, e.g., for key generation or authentication purposes. However, they can be vulnerable to modeling attacks that predict the output for an unknown challenge, based on a set of known challenge/response pairs (CRPs). In addition, an attacker may benefit from power side-channels to break a PUF’s security. Although such attacks have been extensively discussed in literature, the effect of device aging on the efficacy of these attacks is still an open question. Accordingly, in this paper, we focus on the impact of aging on Arbiter-PUFs and one of its modeling-resistant counterparts, the Voltage Transfer Characteristic (VTC) PUF. We present the results of our SPICE simulations used to perform modeling attack via Machine Learning (ML) schemes on the devices aged from 0 to 20 weeks. We show that aging has a significant impact on modeling attacks. Indeed, when the training dataset for ML attack is extracted at a different age than the evaluation dataset, the attack is greatly hindered despite being performed on the same device. We show that the ML attack via power traces is particularly efficient to recover the responses of the anti-modeling VTC PUF, yet the aging still contributes to enhance its security.

Download Paper (PDF; Only available from the DATE venue WiFi)

OFFLINE MODEL GUARD: SECURE AND PRIVATE ML ON MOBILE DEVICES

Speaker: Emmanuel Stapf, TU Darmstadt, DE

Authors: Sebastian P. Bayerl1, Tommaso Frassetto2, Patrick Jauernig2, Korbinian Riedhammer1, Ahmad-Reza Sadeghi1, Thomas Schneider2 and Christian Weier1

1TH Nürnberg, DE; 2TU Darmstadt, DE

Abstract

Performing machine learning tasks in mobile applications yields a challenging conflict of interest: highly sensitive client information (e.g., speech data) should remain private while also the intellectual property of service providers (e.g., model parameters) must be protected. Cryptographic techniques offer secure solutions for this, but have an unacceptable overhead and moreover require frequent network interaction. In this work, we design a practically efficient hardware-based solution. Specifically, we build Offline Model Guard (OMG) to enable privacy-preserving machine learning on the predominant mobile computing platform ARM - even in offline scenarios. By leveraging a trusted execution environment for strict hardware-enforced isolation from other system components, OMG guarantees privacy of client data, secrecy of provided models, and integrity of processing algorithms. Our prototype implementation on an ARM MiKey 960 development board performs privacy-preserving keyword recognition using TensorFlow Lite for Microcontrollers in real time.

Download Paper (PDF; Only available from the DATE venue WiFi)

SECURITY ENHANCEMENT FOR RRAM COMPUTING SYSTEM THROUGH OBfuscating CROSSBAR ROW CONNECTIONS

Speaker: Minhui Zou, Nanjing University of Science and Technology, CN

Authors: Minhui Zou1, Zhenhua Zhu2, Yi Cai2, Junlong Zhou1, Chengliang Wang2 and Yu Wang2

1Nanjing University of Science and Technology, CN; 2Tsinghua University, CN; 3Chongqing University, CN

Abstract

Neural networks (NN) have gained great success in visual object recognition and natural language processing, but this kind of data-intensive applications requires huge data movements between computing units and memory. Emerging resistive random-access memory (RRAM) computing systems have demonstrated great potential in avoiding the huge data movements by performing matrix-vector-multiplications in memory. However, the nonvolatility of the RRAM devices may lead to potential stealing of the NN weights stored in crossbars and the adversary could extract the NN models from the stolen weights. This paper proposes an effective security enhancing method for RRAM computing systems to thwart this sort of piracy attack. We first analyze the theft methods of the NN weights. Then we propose an efficient security enhancing technique based on obfuscating the row connections between positive crossbars and their pairing negative crossbars. Two heuristic techniques are also presented to optimize the hardware overhead of the obfuscation module. Compared with existing NN security work, our method eliminates the additional RRAM writing operations used for encryption/decryption, without shortening the lifetime of RRAM computing systems. The experiment results show that the proposed methods ensure the trial times of brute-force attack are more than $(16)^{17}$ and the classification accuracy of the incorrectly extracted NN models is less than 20%, with minimal area overhead.

Download Paper (PDF; Only available from the DATE venue WiFi)
18:30 4.7.2 MODELING A FLOATING-GATE MEMRISTIVE DEVICE FOR COMPUTER AIDED DESIGN OF NEUROMORPHIC COMPUTING
Speaker: Loai Danial, Technion, IL
Authors: Loai Danial1, Vasu Gupta2, Evgeny Pikhay2, Yakov Rozzin1 and Shahar Kvatinisky 1
Abstract
Memristive technology is still not mature enough for the very large-scale integration necessary to obtain practical value from neuromorphic systems. While nonvolatile floating-gate “synapse transistors” have been implemented in very large-scale integrated neuromorphic systems, their large footprint still constrains an upper bound on the overall performance. A two-terminal floating-gate memristive device can combine the technological maturity of the floating-gate transistor and the conceptual novelty of the memristor using a standard CMOS process. In this paper, we present a top-down computer aided design framework of the floating-gate memristive device and show its potential in neuromorphic computing. Our framework includes a Verilog-A SPICE model, small-signal schematics, a stochastic model, Monte-Carlo simulations, layout, DRC, LVS, and RC extraction.
Download Paper (PDF; Only available from the DATE venue WiFi)

18:00 4.7.3 GROUND PLANE PARTITIONING FOR CURRENT RECYCLING OF SUPERCONDUCTING CIRCUITS
Speaker: Naveen Katam, University of Southern California, US
Authors: Naveen Kumar Katam, Bo Zhang and Massoud Pedram, University of Southern California, US
Abstract
Superconducting single flux quantum (SFQ) technology using Josephson junctions (JJs) is an excellent choice for the computing fabrics of the future. Current recycling is a necessary technique for the implementation of large SFQ circuits with energy-efficiency, where circuit partitions with similar bias current requirements are biased serially. Though this technique has been verified for small scale circuits, it has not been implemented for large circuits as there is no trivial way to partition the circuit into circuit blocks with separate ground planes. The major constraints for partitioning are (1) equal bias current and (2) equal area for all the partitions; (3) minimize the connections between adjacent ground planes with high-cost for non-adjacent planes. For the first time, all these constraints are formulated into a cost function and it is minimized with the gradient descent method. The algorithm takes a circuit netlist and the intended number of partitions as inputs and gives the output as groups of cells belonging to separate ground planes. It minimizes the connections among different ground planes and gives a solution on which the current recycling technique can be implemented. The parameters of cost function have been initialized randomly along with minimizing the dimensions to find the solution quickly. On average, 30% of connections are between non-adjacent ground planes for the given benchmark circuits.
Download Paper (PDF; Only available from the DATE venue WiFi)

18:15 4.7.4 SILICON PHOTONIC MICRORING RESONATORS: DESIGN OPTIMIZATION UNDER FABRICATION NON-UNIFORMITY
Speaker: Mahdi Nikdast, Colorado State University, US
Authors: Asif Mirza, Febin Sunny, Sudeep Pasricha and Mahdi Nikdast, Colorado State University, US
Abstract
Microring resonators (MRRs) are very often considered as the primary building block in silicon photonic integrated circuits (PICs). Despite many advantages, MRRs are considerably sensitive to fabrication non-uniformity (a.k.a. fabrication process variations), necessitating the use of power-hungry compensation methods (e.g., thermal tuning) to guarantee their reliable operation. Moreover, the design space of MRRs is complicated and includes several highly correlated design parameters, preventing designers from easily exploring and optimizing the design of MRRs against fabrication process variations (FPVs). In this paper, for the first time, we present a comprehensive design space exploration and optimization of MRRs against FPVs. In particular, we indicate how physical design parameters in MRRs can be optimized during design time to enhance their tolerance to FPVs while also improving the insertion loss and quality factor in such devices. Fabrication results obtained by measuring multiple fabricated MRRs designed using our design optimization solution demonstrate a significant 70% improvement on average in MRRs tolerance to different FPVs. Such improvement indicates the efficiency of our novel design optimization solution in reducing the tuning power required for reliable operation of MRRs.
Download Paper (PDF; Only available from the DATE venue WiFi)

18:30 IP2-8, 849 CURRENT-MODE CARRY-FREE MULTIPLIER DESIGN USING A MEMRISTOR-TRANSISTOR CROSSBAR ARCHITECTURE
Speaker: Shengqi Yu, Newcastle University, GB
Authors: Shengqi Yu1, Ahmed Soltan2, Rishad Shafi1, Thanasin Bunnam1, Domenico Balsamo1, Fei Xia1 and Alex Yakovlev 1
Abstract
Traditional multipliers consist of complex logic components. They are a major energy and performance contributor of modern compute-intensive applications. As such, designing multipliers with reduced energy and faster speed has remained a challenging task. This paper presents a novel, carry-free multiplier, which is suitable for next-generation of energy-constrained applications. The multiplier circuit consists of an array of memristor-transistor cells that can be selected (i.e., turned ON or OFF) using a combination of DC bias voltages based on the operand values. When a cell is selected it contributes to current in the path, which is then amplified by current mirrors with variable transistor gate sizes. The different current paths are connected to a node for analogously accumulating the currents to produce the multiplier output directly, removing the carry propagation stages, typically seen in traditional digital multipliers. An essential feature of this multiplier is autonomous survivability, i.e., when the power is below this threshold the logic state automatically retains at a zero-cost due to the non-volatile properties of memristors.
Download Paper (PDF; Only available from the DATE venue WiFi)

18:31 IP2-9, 88 N-BIT DATA PARALLEL SPIN WAVE LOGIC GATE
Speaker: Abdulkader Mahmoud, TU Delft, NL
Authors: Abdulkader Mahmoud1, Frederic Vanderveken2, Florin Ciobotaru2, Christoph Adelmann2, Sorin Cotofana1 and Said Hamdioui 1
Abstract
Due to their very nature, Spin Waves (SWs) created in the same waveguide, but with different frequencies, can coexist while selectively interacting with each other. As such, designing spin-wave based processors can lead to high performance with low power overhead. However, in ordered Magnetic Field (OMF) based spin-wave processors, the current recycling technique is essential to achieve high energy efficiency. In this work, we present a novel 8-bit parallel data-parallel spin-wave-based logical gate implementation. We demonstrate the feasibility of our approach using 8-bit data parallel gate with equivalent SW gate based implementation. Our evaluation indicates that 8-bit data 3-input Majority gate implementation requires 4.16x less area than the scalar SW gate based equivalent counterpart while preserving the same delay and energy consumption figures.
Download Paper (PDF; Only available from the DATE venue WiFi)

18:30 End of session
## UB04 Session 4

**Date:** Tuesday 10 March 2020  
**Time:** 17:30 - 19:30  
**Location / Room:** Booth 11, Exhibition Area

<table>
<thead>
<tr>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
</table>
| UB04.1  | FLETCHER: TRANSPARENT GENERATION OF HARDWARE INTERFACES FOR ACCELERATING BIG DATA APPLICATIONS | Zaid Al-Ars, Johan Peletenburg, Jeroen van Straten, Matthijs Brobel and Joost Hoozemans, TU Delft, NL  
**Abstract**  
This demo created by TUDelft is a software-hardware framework to allow for an efficient integration of FPGA hardware accelerators both on edge devices as well as in the cloud. The framework is called Fletcher, which is used to automatically generate data communication interfaces in hardware based on the widely used big data format Apache Arrow. This provides two distinct advantages. On the one hand, since the accelerators use the same data format as the software, data communication bottlenecks can be reduced. On the other hand, since a standardized data format is used, this allows for easy-to-use interfaces on the accelerator side, thereby reducing the design and development time. The demo shows how to use Fletcher for big data acceleration to deconstruct Snappy compressed files and perform filtering on the whole Wikipedia body of text. The demo enables 25 GB/s processing throughput.  
More information ...

| UB04.2  | ELSA: EIGENVALUE BASED HYBRID LINEAR SYSTEM ABSTRACTION: BEHAVIORAL MODELING OF TRANSISTOR-LEVEL CIRCUITS USING AUTOMATIC ABSTRACTION TO HYBRID AUTOMATA | Ahmad Taraf and Lars Hedrich, University of Frankfurt, DE  
**Abstract**  
Model abstraction of transistor-level circuits, while preserving an accurate behavior, is still an open problem. In this demo an approach is presented that automatically generates a hybrid automaton (HA) with linear states from an existing circuit netlist. The approach starts with a netlist at transistor level with full SPICE accuracy and ends at the system level description of the circuit in matlab or in Verilog-A. The resulting hybrid automaton exhibits linear behavior as well as the technology dependent nonlinear e.g. limiting behavior. The accuracy and speed-up of the Verilog-A generated models is evaluated based on several transistor level circuit abstractions of simple operational amplifiers up to a complex filters. Moreover, we verify the equivalence between the generated model and the original circuit. For the generated models in matlab syntax, a reachability analysis is performed using the reachability tool cora.  
More information ...

| UB04.3  | SRSN: SECURE RECONFIGURABLE TEST NETWORK | Vincent Reynaud,1 Emanuele Valea,2 Paolo Maistri,1 Regis Laveugle,1 Marie-Lise Flottes,2 Sophie Dupuis,2 Bruno Pouze,2 and Giorgio Di Natale1  
1TIMA Laboratory, FR; 2LIRMM, FR  
**Abstract**  
The critical importance of testability for electronic devices led to the development of IEEE test standards. These methods, if not protected, offer a security backdoor to attackers. This demonstrator illustrates a state-of-the-art solution that prevents unauthorized usage of the test infrastructure based on the IEEE 1687 standard and implemented on an FPGA target.  
More information ...

| UB04.4  | LAGARTO: FIRST SILICON RISC-V ACADEMIC PROCESSOR DEVELOPED IN SPAIN | Guillem Cabo Pitarch1, Cristobal Ramirez Lazo1, Julian Pavon Rivera1, Vasilis Kostalabros1, Carlos Rojas Morales1, Miguel Moreto1, Jaume Abella1, Francisco J. Calvo1, Adrian Cristal1, Roger Figueras1, Alberto Gonzalez1, Carles Hernandez1, Cesar Hernandez1, Nei Leyva2, Joan Marin1, Ricardo Martinez1, Jonatan Mendozal, Francesc Moll1, Marco Antonio Ramirez1, Carlos Rojas1, Antonio Rubio1, Abraham Ruiz1, Nehir Sonmez1, Lluis Teres1, Osman Unsal1, Mateo Valero1, Ivan Vargas1 and Luis Villa2  
1BSC / UPC, ES; 2CIC-IPN, MX; 3IMB-CNM (CSIC), ES; 4UPC, ES; 5BSC, ES  
**Abstract**  
Open hardware is a possibility that has emerged in recent years and has the potential to be as disruptive as Linux was once, an open source software paradigm. If Linux managed to lessen the dependence of users in large companies providing software and software applications, it is envisioned that hardware based on ISAs open source can do the same in their own field. In the Lagarto tapeout four research institutions were involved: Centro de Investigación en Computación of the Mexican IPN, Centro Nacional de Microelectrónica of the CSIC, Universitat Politècnica de Catalunya (UPC) and Barcelona Supercomputing Center (BSC). As a result, many bachelor, master and PhD students had the chance to achieve real-world experience with ASIC design and achieve a functional SoC. In the booth, you will find a live demo of the first ASIC and prototypes running on FPGA of the next versions of the SoC and core.  
More information ...

| UB04.5  | LEARNV: LEARNV: A RISC-V BASED EMBEDDED SYSTEM DESIGN FRAMEWORK FOR EDUCATION AND RESEARCH DEVELOPMENT | Nouredine Ait Saïd and Mourin Benabenbeni, TIMA Laboratory, FR  
**Abstract**  
Designing a modern System on a Chip is based on the joint design of hardware and software (co-design). However, understanding the tight relationship between hardware and software is not straightforward. Moreover to validate new concepts in SoC design from the idea to the hardware implementation is time-consuming and often slowed by legacy issues (intellectual property of hardware blocks and expensive commercial tools). To overcome these issues we propose to use the open-source Rocket Chip environment for educational purposes, combined with the open-source LowRisc architecture to implement a custom SoC design on an FPGA board. The demonstration will present how students and engineers can take benefit from the environment to deepen their knowledge in HW and SW co-design. Using the LowRisc architecture, an image classification application based on the use of CNNs will serve as a demonstrator of the whole open-source hardware and software flow and will be mapped on a Nexys A7 FPGA board.  
More information ...

| UB04.6  | CSI-REPUTE: A LOW POWER EMBEDDED DEVICE CLUSTERING APPROACH TO GENOME READ MAPPING | Tousif Rahman1, Sidharth Maheshwari1, Rishad Shaik1, Ian Wilson1, Alex Yakovlev1 and Amit Acharyya2  
1Newcastle University, GB; 2TIIT Hyderabad, IN  
**Abstract**  
The big data challenge of genomics is rooted in its requirements of extensive computational capability and results in large power and energy consumption. To encourage widespread usage of genome assembly tools there must be a transition from the existing predominantly software-based mapping tools, optimized for homogeneous high-performance systems, to more heterogeneous low power and cost-effective mapping systems. This demonstration will show a cluster system implementation for the REPUTE algorithm, (An OpenCL based Read Mapping Tool for Embedded Genomics) where cluster nodes are composed of low power single board computer (SBC) devices and the algorithm is deployed on each node spreading the genomic workload. We propose a working concept prototype to challenge current conventional high-performance many-core CPU based cluster nodes. This demonstration will highlight the advantage in the power and energy domains of using SBC clusters enabling potential solutions to low-cost genomics.  
More information ...
Abstract
Graphics processing units (GPUs) can provide the increased performance required in future critical systems, i.e. automotive and avionics. However, their programming models, e.g. CUDA or OpenCL, cannot be used in such systems as they violate safety critical programming guidelines. Brook SC (https://github.com/kosmid/brook) was developed in UPC/BSC to allow safety-critical applications to be programmed in a CUDA-like GPU language, Brook, which enables the certification while increasing productivity. In our demo, an avionics application running on a realistic safety critical GPU software stack and hardware is shown. Credited in this Bachelor’s thesis project, which was awarded a 2019 HiPEAC Technology Transfer Award, an Airbus prototype application performing general-purpose computations with a safety-critical graphics API was ported to Brook SC in record time, achieving an order of magnitude reduction in the lines of code to implement the same functionality without performance penalty.

More information ...

Brook SC: High-level Certification-Friendly Programming for GPU-Powered Safety Critical Systems
Authors:
Marc Benito, Matina Maria Trompouki and Leonidas Kosmidis, BSC / UPC, ES

Abstract
Distributing time-sensitive applications on edge computing environments.
Authors:
Eudald Sabaté Creixell 1, Unai Perez Mendizabal 1, Elli Kartsakli 2, Maria A. Serrano Gracia 3 and Eduardo Quiliones Moreno 3
1IBSC / UPC, ES; 2IBSC, GR; 3IBSC, ES

Abstract
The proposed demonstration aims to showcase the capabilities of a task-based distributed programming framework for the execution of real-time applications in edge computing scenarios, in the context of smart cities. Edge computing shifts the computation close to the data source, alleviating the pressure on the cloud and reducing application response times. However, the development and deployment of distributed real-time applications is complex, due to the heterogeneous and dynamic edge environment where resources may not always be available. To address these challenges, our demo employs COMPSS, a highly portable and infrastructure-agnostic programming model, to efficiently distribute time-sensitive applications across the compute continuum. We will exhibit how COMPSS distributes the workload on different edge devices (e.g., NVIDIA GPUs and a Raspberry Pi), and how COMPSS re-adapts this distribution upon the availability (connection or disconnection) of devices.

More information ...

5.1 Special Day on “Embedded AI”: Tutorial Overviews

Date: Wednesday 11 March 2020
Time: 08:30 - 10:00
Location / Room: Amphithéâtre Jean Prouve
Chair: Dmitri Strukov, University of California, Santa Barbara, US
Co-Chair: Bernabe Linares-Barranco, CSIC, ES

This session aims to provide a more tutorial overview of hardware AI case studies and some proposed solutions, problems, and challenges.

Time Label Presentation Title Authors
08:30 5.1.1 NEURAL NETWORKS CIRCUITS BASED ON RESISTIVE MEMORIES
Author: Carlo Reita, CEA, FR
Abstract
In recent years, the field of Neural Networks has found a new golden age after nearly twenty years of lessened interest. Under the heading of Artificial Intelligence (AI) a large number of Deep Neural Networks (DNNs) have recently found application in image processing, management of information in large databases, decision aids, natural language recognition, etc. Most of these applications rely on algorithms that run on standard computing systems and sometimes make use of specific accelerators like Graphic Processor Units (GPUs) or dedicated highly parallel processors. In effect, a common operation in all NN algorithms is the scalar product of two vectors and its optimisation is of paramount importance to reduce computational time and energy. In particular, the energy element is relevant for all embedded applications that cannot rely on cooling and/or unlimited power supply. The availability of resistive memories, with their unique capability of both storing computational values and of performing analog multiplication by the use of ohm’s law, allows new circuit architectures where the latency, bandwidth limitations and power consumption issues associated to the use of conventional SRAM, DRAM and Flash memories can be greatly improved upon. In the presentation, some examples of advantageous use of resistive memories in NN circuits will be shown and some of their peculiarities will be discussed.
5.2 Machine Learning Approaches to Analog Design

Date: Wednesday 11 March 2020
Time: 08:30 - 10:00
Location / Room: Chamrousse

Chair: Marie-Minerve Louerat, Sorbonne University Lip6, FR
Co-Chair: Sebastien Cliquennois, STMicroelectronics, FR

This session presents recent advances in machine learning approaches to support the design of analog and mixed-signal circuits. Techniques such as reinforced learning and convolutional networks are employed to address circuit and layout optimization. The presented techniques have a great potential for seeding innovative solutions to face current and future challenges in this field.

08:30 | 5.2.1 | AUTOCKT: DEEP REINFORCEMENT LEARNING OF ANALOG CIRCUIT DESIGNS
Speaker: Keertana Settaluri, University of California, Berkeley, US
Authors: Keertana Settaluri, Ameer Haj-Ali, Qijing Huang, Kouros Hakkamashesh and Borivoje Nikolic, University of California, Berkeley, US

Abstract
The need for domain specialization under energy constraints in deeply-scaled CMOS has been driving the need for agile development of Systems on a Chip (SoCs). While digital subsystems have design flows that are conducive to rapid iterations from specification to layout, analog and mixed-signal modules face the challenge of a long human-in-the-middle iteration loop that requires expert intuition to verify that post-layout circuit parameters meet the original design specification. Existing automated solutions that optimize circuit parameters for a given target design specification have limitations of being schematic-only, inaccurate, sample-inefficient or not generalizable. This work presents AutoCkt, a deep-reinforcement learning tool that not only finds post-layout circuit parameters for a given target specification, but also gains knowledge about the entire design space through a sparse subsampling technique. Our results show that for multiple circuit topologies, the trained AutoCkt agent is able to converge and meet all target specifications on at least 96.3% of tested design goals in schematic simulation, on average 40X faster than a traditional genetic algorithm. Using the Berkeley Analog Generator, AutoCkt is able to design 40 SoCs passed operational amplifiers in 68 hours, 9.6X faster than the state-of-the-art when considering layout parasitics.

Download Paper (PDF; Only available from the DATE venue WiFi)

09:00 | 5.2.2 | TOWARDS DECRYPTING THE ART OF ANALOG LAYOUT: PLACEMENT QUALITY PREDICTION VIA TRANSFER LEARNING
Speaker: David Pan, University of Texas at Austin, US
Authors: Mingjie Liu, Keren Zhu, Jiaqi Gu, Linxiao Shen, Xiuyan Tang, Nan Sun and David Z. Pan, University of Texas at Austin, US

Abstract
Despite tremendous efforts in analog layout automation, little adoption has been demonstrated in practical design flows. Traditional analog layout synthesis tools use various heuristic constraints to prune the design space to ensure post layout performance. However, these approaches provide limited guarantee and poor generalizability due to a lack of model mapping layout properties to circuit performance. In this paper, we attempt to shorten the gap in post-layout performance modeling for analog circuits with a quantitative statistical approach. We leverage a state-of-the-art automatic layout tool and industry-level simulator to generate labeled training data in an automatic manner. We propose a 3D convolutional neural network (CNN) model to predict the relative placement quality using well-crafted placement features. To achieve data-efficiency for practical usage, we further propose a transfer learning scheme that greatly reduces the amount of data needed. Our model would enable early pruning and efficient design explorations for practical layout design flows. Experimental results demonstrate the effectiveness and generalizability of our method across different operational transconductance amplifier (OTA) designs.

Download Paper (PDF; Only available from the DATE venue WiFi)

09:30 | 5.2.3 | DESIGN OF MULTI-OUTPUT SWITCHED-CAPACITOR VOLTAGE REGULATOR VIA MACHINE LEARNING
Speaker: Zhiyuan Zhou, Washington State University, US
Authors: Zhiyuan Zhou1, Syrine Belakaria2, Aryan Deshwal2, Wookpyo Hong1, Jana Doppa2, Partha Pratim Pande 1 and Deukhyoun Heo 1
1Washington State University, US; 2Washington State University, US

Abstract
Efficiency of power management system (PMS) is one of the key performance metrics for highly integrated system on chips (SoCs). Towards the goal of improving power efficiency of SoCs, we make two key technical contributions in this paper. First, we develop a multi-output switched-capacitor voltage regulator (SCVR) with a new flying capacitor crossing technique (FCCT) and cloud-capacitor method. Second, to optimize the design parameters of SCVR, we introduce a novel machine-learning (ML)-inspired optimization framework to reduce the number of expensive design simulations. Simulation shows that power loss of the multi-output SCVR with FCCT is reduced by more than 40% compared to conventional multiple single-output SCVRs. Our ML-based design optimization framework is able to achieve more than 90% reduction in the number of simulations needed to uncover optimized circuit parameters of the proposed SCVR.
centric attacks and against software-driven attacks on hardware. It brings together researchers and industry practitioners who deal with secure composition: security-oriented from a purely hardware-centric and from a hardware-software perspective in a more complex system. It will also target composition of countermeasures against hardware.

Physical and side-channel attacks add another level of complexity to the problem of secure composition. Moreover, practical hardware systems include software limited to software systems, and unique security challenges arise when a real system, composed of a range of hardware components with different owners and trust assumptions and techniques that can facilitate composition and integration of hardware systems and achieve security guarantees. Theoretical foundations of secure composition are currently coupled with safety-critical sub-systems (driving assistance) and security-sensitive sub-systems such as online payment and others. Moreover, a system's hardware components are now often directly accessible to the end users and thus vulnerable to physical attacks. The goal of this hot-topic session is to establish a common understanding of principles and larger. From a certain imager size, the proposed simulation method yields simulation times two order of magnitude faster than an implicit method based on Newton-Raphson iterations. However, given that explicit methods do not require the computation of time-consuming matrix factorizations, the overall simulation time is reduced. The technique described in this work has been implemented on a NVIDIA general purpose GPU and has been tested simulating the Gaussian filtering operation performed by a smart CMOS image sensor. Such devices are used to perform computation on the edge and include built-in image processing functions. Among those, the Gaussian filtering is one of the most common functions, since it is a basic task for early vision processing. These smart sensors are increasingly complex and hence the time required to simulate them during their design cycle is also larger and larger.

Modern electronic systems become evermore complex, yet remain modular, with integrated circuits (ICs) acting as versatile hardware components at their heart. Electronic design automation (EDA) for ICs has focused traditionally on power, performance, and area. However, given the rise of hardware-centric security threats, we believe that EDA must also adopt related notions like secure by design and secure composition of hardware. Despite various promising studies, we argue that some aspects still require more efforts, for example: effective means for compilation of assumptions and constraints for security schemes, all the way from the system level down to the “bare metal”; modeling, evaluation, and consideration of security-relevant metrics; or automated and holistic synthesis of various countermeasures, without inducing negative cross-effects. In this paper, we first introduce hardware security for the EDA community. Next we review prior (academic) art for EDA-driven security evaluation and implementation of countermeasures. We then discuss strategies and challenges for advancing research and development toward secure composition of circuits and systems.

**References**

1. Tom Kozmienski, University of Southampton, GB
2. Gines Domenech-Aaens and Tom J Kozmienski
3. Universidad Politecnica de Cartagena, ES
4. University of Southampton, GB
5. Ilia Polian, Stuttgart University, DE
6. Francesco Regazzoni, ALARI, CH
7. Tim Güneysu, Ruhr-University Bochum, DE
8. Debdeep Mukhopadhyay, IIT Kharagpur, IN
9. Annelie Heuser, ALaRI, CH
10. Dey Soumyajit, Boston College, US
11. Yunsi Fei, University of Sheffield, GB
12. Tim Güneysu, University of Sheffield, GB
13. Ila Polian, Stuttgart University, DE
14. Elif Bilge Kavun, Bar-Ilan University, IL
15. Anupam Chattopadhyay, IIT Kharagpur, IN
16. Yunsi Fei, Ruhr-University Bochum, DE
17. Anand Prakash, IBM

**Abstract**

Human Body Communication has shown great promise to replace wireless communication for information exchange between wearable devices of a body area network. However, there are very few studies in literature, that systematically study the channel loss of capacitive HBC for wearable devices over a wide-frequency range with different terminations at the receiver, partly due to the need for miniaturized wearable devices for an accurate study. This paper, for the first time, measures the channel loss of capacitive HBC from 100kHz to 1GHz for both high-impedance and 50 ohm terminations using wearable, battery powered devices which is mandatory for accurate measurement of the HBC channel-loss, due to ground coupling effects. Results show that high impedance termination leads to a significantly lower channel loss (40 dB improvement at 1MHz), as compared to 50 ohm termination at low frequencies. This difference steadily decreases with increasing frequency, until they become similar near 80MHz. Beyond 100MHz inter-device coupling dominates, thereby preventing accurate measurements of channel loss of the human body. The measured results provide a consistent wearable, wide-frequency HBC channel loss data and could serve as a backbone for the emerging field of HBC by aiding in the selection of an appropriate operation frequency and termination.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**Title:** TOWARDS SECURE COMPOSITION OF INTEGRATED CIRCuits and ELECTRONic SYSTEMS: ON THE ROLE OF EDA

**Speaker:** Johann Knechtel, New York University Abu Dhabi, AE

**Authors:** Johann Knechtel, Elif Bilge Kayuan, Francesco Regazzoni, Annelie Heuser, Anupam Chattopadhyay, Debdeep Mukhopadhyay, Dey Soumyajit, Yunsi Fei, Yaakov Belenky, Itamar Levi, Tim Güneysu, Patrick Schaumüller, Ilia Polian

**University:** New York University Abu Dhabi, AE

**Abstract:** Modern electronic systems become evermore complex, yet remain modular, with integrated circuits (ICs) acting as versatile hardware components at their heart. Electronic design automation (EDA) for ICs has focused traditionally on power, performance, and area. However, given the rise of hardware-centric security threats, we believe that EDA must also adopt related notions like secure by design and secure composition of hardware. Despite various promising studies, we argue that some aspects still require more efforts, for example: effective means for compilation of assumptions and constraints for security schemes, all the way from the system level down to the “bare metal”; modeling, evaluation, and consideration of security-relevant metrics; or automated and holistic synthesis of various countermeasures, without inducing negative cross-effects. In this paper, we first introduce hardware security for the EDA community. Next we review prior (academic) art for EDA-driven security evaluation and implementation of countermeasures. We then discuss strategies and challenges for advancing research and development toward secure composition of circuits and systems.
| Time   | Label | Presentation Title                                                                 | Authors                                                                                                                                                                                                 | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors                                      | Authors catalogue |
The session explores the use of state-of-the-art model-based analysis and verification techniques to secure and improve the performance of embedded systems. More specifically, Todd Austin, University of Michigan, US

Co-Chair: Ylies Falcone, University Grenoble Alpes and Inria, FR

Chair: Location / Room: 5.5 Model-Based Analysis and Security

Abstract
Multicore chips are expected to rely on coherent shared memory. Albeit the coherence hardware can scale gracefully, the protocol state space grows exponentially with core count. That is why design verification requires directed test generation (DTG) for dynamic coverage control under the tight time constraints resulting from slow simulation and short verification budgets. Next generation EDA tools are expected to exploit Machine Learning for reaching high coverage in less time. We propose a technique that addresses DTG as a decision process and tries to find a decision-making policy for maximizing the cumulative coverage, as a result of successive actions taken by an agent. Instead of simply relying on learning, our technique builds upon the legacy from constrained random test generation (RTG). It casts DTG as coverage-driven RTG, and it explores distinct RTG engines subject to progressively tighter constraints. We compared three Reinforcement Learning generators with a state-of-the-art generator based on Genetic Programming. The experimental results show that the proper enforcement of constraints is more efficient for guiding learning towards higher coverage than simply letting the generator learn how to select the most promising memory events for increasing coverage. For a 3-level MESI 32-core design, the proposed approach led to the highest observed coverage (95.81%), and it was 2.4 times faster than the baseline generator to reach the latter’s maximal coverage.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

Abstract
Formal verification methods have made huge progress over the last decades. However, proving the correctness of arithmetic circuits involving integer multipliers still drives the verification techniques to their limits. Recently, Symbolic Computer Algebra (SCA) methods have shown good results in the verification of both large and non-trivial multipliers. Their success is mainly based on (1) reverse engineering and identifying basic building blocks, (2) finding converging gate cones which start from the basic building blocks and (3) early removal of redundant terms (vanishing monomials) to avoid the blow-up during backward rewriting. Despite these important accomplishments, verifying optimized and technology-mapped multipliers is an almost unexplored area.

This creates major barriers for industrial use as most of the designs are area and delay optimized. To overcome the barriers, we propose a novel SCA-method which supports the formal verification of a large variety of optimized multipliers. Our method takes advantage of a dynamic substitution ordering to avoid the monomial explosion during backward rewriting. Experimental results confirm the efficiency of our approach in the verification of a wide range of optimized multipliers including industrial benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

Abstract
Formal verification methods have made huge progress over the last decades. However, proving the correctness of arithmetic circuits involving integer multipliers still drives the verification techniques to their limits. Recently, Symbolic Computer Algebra (SCA) methods have shown good results in the verification of both large and non-trivial multipliers. Their success is mainly based on (1) reverse engineering and identifying basic building blocks, (2) finding converging gate cones which start from the basic building blocks and (3) early removal of redundant terms (vanishing monomials) to avoid the blow-up during backward rewriting. Despite these important accomplishments, verifying optimized and technology-mapped multipliers is an almost unexplored area. This creates major barriers for industrial use as most of the designs are area and delay optimized. To overcome the barriers, we propose a novel SCA-method which supports the formal verification of a large variety of optimized multipliers. Our method takes advantage of a dynamic substitution ordering to avoid the monomial explosion during backward rewriting. Experimental results confirm the efficiency of our approach in the verification of a wide range of optimized multipliers including industrial benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

Abstract
Formal verification methods have made huge progress over the last decades. However, proving the correctness of arithmetic circuits involving integer multipliers still drives the verification techniques to their limits. Recently, Symbolic Computer Algebra (SCA) methods have shown good results in the verification of both large and non-trivial multipliers. Their success is mainly based on (1) reverse engineering and identifying basic building blocks, (2) finding converging gate cones which start from the basic building blocks and (3) early removal of redundant terms (vanishing monomials) to avoid the blow-up during backward rewriting. Despite these important accomplishments, verifying optimized and technology-mapped multipliers is an almost unexplored area.

This creates major barriers for industrial use as most of the designs are area and delay optimized. To overcome the barriers, we propose a novel SCA-method which supports the formal verification of a large variety of optimized multipliers. Our method takes advantage of a dynamic substitution ordering to avoid the monomial explosion during backward rewriting. Experimental results confirm the efficiency of our approach in the verification of a wide range of optimized multipliers including industrial benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

Abstract
Formal verification methods have made huge progress over the last decades. However, proving the correctness of arithmetic circuits involving integer multipliers still drives the verification techniques to their limits. Recently, Symbolic Computer Algebra (SCA) methods have shown good results in the verification of both large and non-trivial multipliers. Their success is mainly based on (1) reverse engineering and identifying basic building blocks, (2) finding converging gate cones which start from the basic building blocks and (3) early removal of redundant terms (vanishing monomials) to avoid the blow-up during backward rewriting. Despite these important accomplishments, verifying optimized and technology-mapped multipliers is an almost unexplored area.

This creates major barriers for industrial use as most of the designs are area and delay optimized. To overcome the barriers, we propose a novel SCA-method which supports the formal verification of a large variety of optimized multipliers. Our method takes advantage of a dynamic substitution ordering to avoid the monomial explosion during backward rewriting. Experimental results confirm the efficiency of our approach in the verification of a wide range of optimized multipliers including industrial benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

Abstract
Formal verification methods have made huge progress over the last decades. However, proving the correctness of arithmetic circuits involving integer multipliers still drives the verification techniques to their limits. Recently, Symbolic Computer Algebra (SCA) methods have shown good results in the verification of both large and non-trivial multipliers. Their success is mainly based on (1) reverse engineering and identifying basic building blocks, (2) finding converging gate cones which start from the basic building blocks and (3) early removal of redundant terms (vanishing monomials) to avoid the blow-up during backward rewriting. Despite these important accomplishments, verifying optimized and technology-mapped multipliers is an almost unexplored area.

This creates major barriers for industrial use as most of the designs are area and delay optimized. To overcome the barriers, we propose a novel SCA-method which supports the formal verification of a large variety of optimized multipliers. Our method takes advantage of a dynamic substitution ordering to avoid the monomial explosion during backward rewriting. Experimental results confirm the efficiency of our approach in the verification of a wide range of optimized multipliers including industrial benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

Abstract
Formal verification methods have made huge progress over the last decades. However, proving the correctness of arithmetic circuits involving integer multipliers still drives the verification techniques to their limits. Recently, Symbolic Computer Algebra (SCA) methods have shown good results in the verification of both large and non-trivial multipliers. Their success is mainly based on (1) reverse engineering and identifying basic building blocks, (2) finding converging gate cones which start from the basic building blocks and (3) early removal of redundant terms (vanishing monomials) to avoid the blow-up during backward rewriting. Despite these important accomplishments, verifying optimized and technology-mapped multipliers is an almost unexplored area.

This creates major barriers for industrial use as most of the designs are area and delay optimized. To overcome the barriers, we propose a novel SCA-method which supports the formal verification of a large variety of optimized multipliers. Our method takes advantage of a dynamic substitution ordering to avoid the monomial explosion during backward rewriting. Experimental results confirm the efficiency of our approach in the verification of a wide range of optimized multipliers including industrial benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

Abstract
Formal verification methods have made huge progress over the last decades. However, proving the correctness of arithmetic circuits involving integer multipliers still drives the verification techniques to their limits. Recently, Symbolic Computer Algebra (SCA) methods have shown good results in the verification of both large and non-trivial multipliers. Their success is mainly based on (1) reverse engineering and identifying basic building blocks, (2) finding converging gate cones which start from the basic building blocks and (3) early removal of redundant terms (vanishing monomials) to avoid the blow-up during backward rewriting. Despite these important accomplishments, verifying optimized and technology-mapped multipliers is an almost unexplored area.

This creates major barriers for industrial use as most of the designs are area and delay optimized. To overcome the barriers, we propose a novel SCA-method which supports the formal verification of a large variety of optimized multipliers. Our method takes advantage of a dynamic substitution ordering to avoid the monomial explosion during backward rewriting. Experimental results confirm the efficiency of our approach in the verification of a wide range of optimized multipliers including industrial benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

Abstract
Formal verification methods have made huge progress over the last decades. However, proving the correctness of arithmetic circuits involving integer multipliers still drives the verification techniques to their limits. Recently, Symbolic Computer Algebra (SCA) methods have shown good results in the verification of both large and non-trivial multipliers. Their success is mainly based on (1) reverse engineering and identifying basic building blocks, (2) finding converging gate cones which start from the basic building blocks and (3) early removal of redundant terms (vanishing monomials) to avoid the blow-up during backward rewriting. Despite these important accomplishments, verifying optimized and technology-mapped multipliers is an almost unexplored area.

This creates major barriers for industrial use as most of the designs are area and delay optimized. To overcome the barriers, we propose a novel SCA-method which supports the formal verification of a large variety of optimized multipliers. Our method takes advantage of a dynamic substitution ordering to avoid the monomial explosion during backward rewriting. Experimental results confirm the efficiency of our approach in the verification of a wide range of optimized multipliers including industrial benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)
**IS REGISTER TRANSFER LEVEL LOCKING SECURE?**

**Authors:**
Chandan Karfa, IIT Guwahati, IN

**Abstract**
Register Transfer Level (RTL) locking seeks to prevent intellectual property (IP) theft of a design by locking the RTL description that functions correctly on the application of a key. This paper evaluates the security of a state-of-the-art RTL locking scheme using a satisfiability modulo theories (SMT) based algorithm to retrieve the secret key. The attack first obtains the high-level behavior of the locked RTL, and then uses an SMT based formulation to find so-called distinguishing input patterns (DIPs), i.e., inputs that help eliminate incorrect keys from the keyspace. The attack methodology has two main advantages over the gate-level attacks. First, since the attack handles the design at the RTL, the method scales to large designs. Second, the attack does not apply separate unlocking strategies for the combinational and sequential parts of design; it handles both styles via a unifying abstraction. We demonstrate the attack on locked RTL generated by TAO, a state-of-the-art RTL locking solution. Empirical results show that we can partially or completely break designs locked by TAO.

Download Paper (PDF; Only available from the DATE venue WiFi)

**DESIGN SPACE EXPLORATION FOR MODEL-BASED COMMUNICATION SYSTEMS**

**Authors:**
Valentina Richthammer, University of Ulm, DE

**Abstract**
A main challenge of modern design lies in selecting a suitable combination of subsystems (e.g. Analog Digital/Digital Analog Converters (ADC/DAC), (de)modulators, scramblers, interleavers, and coding and filtering modules) - each of which can be implemented in a multitude of ways. At the same time, the complete modem configuration needs to be tailored to the specific requirements of the intended communication channel or scenario. Therefore, model-based design methodologies have recently been popularized in this field, since their application facilitates the specification of individual modem components that are easily exchanged during the automated synthesisation of the modem. However, this development has resulted in a tremendous increase in the number of synthesizable modem options. In fact, the optimal modem configuration for a communication scenario can not readily be determined, since an exhaustive analysis of all configuration possibilities is computationally intractable. To remedy this, we propose a fully automated Design Space Exploration (DSE) methodology for model-based modem design that combines (I) the metaheuristic exploration and optimization of modem-configuration possibilities with (II) a simulation analysis of suitable measures of communication quality. The presented case study for an acoustic underwater communication scenario supports the described need for novel, automated methodologies in the area of model-based design, since the modem configurations discovered during a comparably short DSE are demonstrated to significantly outperform state-of-the-art modems from literature.

Download Paper (PDF; Only available from the DATE venue WiFi)

**STATISTICAL TIME-BASED INTRUSION DETECTION IN EMBEDDED SYSTEMS**

**Authors:**
Nadir Carreon Rascon, University of Arizona, US

**Abstract**
This paper presents a statistical method based on cumulative distribution functions (CDF) to analyze an embedded system’s behavior to detect anomalous and malicious executions behaviors. The proposed method analyzes the internal timing of the system by monitoring individual operations and sequences of operations, wherein the timing of operations is decomposed into multiple timing subcomponents. Creating the normal model of the system utilizing the internal timing adds resilience to zero-day attacks, and mimics malware. The combination of CDF-based statistical analysis and timing subcomponents enable both higher detection rates and lower false positives rates. We demonstrate the effectiveness of the approach and compare to several state-of-the-art malware detection methods using two embedded systems benchmarks, namely a network connected pacemaker and an unmanned aerial vehicle, utilizing seven different malware.

Download Paper (PDF; Only available from the DATE venue WiFi)

**IFFSET: IN-FIELD FUZZING OF INDUSTRIAL CONTROL SYSTEMS USING SYSTEM EMULATION**

**Authors:**
Dimitrios Tychalas, New York University, US

**Abstract**
Industrial Control Systems (ICS) have evolved in the last decade, shifting from proprietary software/hardware to contemporary embedded architectures paired with open-source operating systems. In contrast to the IT world, where continuous updates and patches are expected, decommissioning always-on ICS for security assessment can incur prohibitive costs to their owner. Thus, a solution for routinely assessing the cybersecurity posture of diverse ICS without affecting their operation is essential. Therefore, in this paper we introduce IFFSET, a platform that leverages full system emulation of Linux-based ICS firmware and utilizes fuzzing for security evaluation. Our platform extracts the file system and kernel information from a live ICS device, building an image which is emulated on a desktop system through QEMU. We employ fuzzing as a security assessment tool to analyze ICS specific libraries and find potential security threatening conditions. We test our platform with commercial PLCs, showcasing potential threats with no interruption to the control process.

Download Paper (PDF; Only available from the DATE venue WiFi)
A logic synthesis toolbox for reducing the multiplicative complexity in logic networks

**Abstract**

Logic synthesis is a fundamental step in the realization of modern integrated circuits. It has traditionally been employed for the optimization of CMOS-based designs, as well as for emerging technologies and quantum computing. Recently, it found application in minimizing the number of AND gates in cryptography benchmarks represented as xor-and graphs (XAGs). The number of AND gates in an XAG, which is called the logic network’s multiplicative complexity, plays a critical role in various cryptography and security protocols such as fully homomorphic encryption (FHE) and secure multi-party computation (MPC). Further, the number of AND gates is also important to assess the degree of vulnerability of a Boolean function, and influences the cost of techniques to protect against side-channel attacks. However, so far a complete logic synthesis flow for reducing the multiplicative complexity in logic networks did not exist or relied heavily on manual manipulations. In this paper, we present a logic synthesis toolbox for cryptography and security applications. The proposed tool consists of powerful transformations, namely resubstitution, refactoring, and rewriting, specifically designed to minimize the multiplicative complexity of an XAG. Our flow is fully automatic and achieves significant results over both EPFL benchmarks and cryptography circuits. We improve the best-known results for cryptography up to 59%, resulting in a normalized geometric mean of 0.82.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

09:00 5.6.2 SAVING POWER BY CONVERTING FLIP-FLOP TO 3-PHASE LATCH-BASED DESIGNS

**Speaker:**
Peter Beerel, University of Southern California, US

**Authors:**
Huimei Cheng, Xi Li, Yichen Gu and Peter Beerel, University of Southern California, US

**Abstract**

Latches are smaller and lower power than flip-flops (FFs) and are typically used in a time-borrowing master-slave configuration. This paper presents an automatic flow for converting arbitrarily-complex single-clock-domain FF-based RTL designs to efficient 3-phase latch-based designs with reduced number of required latches, saving both register and clock-tree power. Post place-and-route results demonstrate that our 3-phase latch-based designs save an average of 15.5% and 18.5% power on a variety of ISCAS, CEP, and CPU benchmark circuits, compared to their more traditional FF and master-slave based alternatives.

**Download Paper (PDF; Only available from the DATE venue WiFi)**
5.7 Stochastic Computing

Time  | Label  | Presentation Title                                                                 | Authors                                                                 |
---    | ---    | ---                                                                                | ---                                                                     |
09:30  | 5.6.3  | COMPUTING THE FULL QUOTIENT IN BI-DECOMPOSITION BY APPROXIMATION                    | Valentina Ciriani, University of Milan, IT                              |
      |        | Speaker: Valentina Ciriani, University of Milan, IT                                |                                                                        |
      |        | Authors: Anna Bernasconi¹, Valentina Ciriani², Jordi Cortadella³ and Tiziano Villa⁴ |                                                                        |
      |        | ¹Università di Pisa, IT; ²Università degli Studi di Milano, IT; ³UPC, ES; ⁴Università di Verona, IT |                                                                        |
      |        | Abstract                                                                             |                                                                         |
      |        | Bi-decomposition is a design technique widely used to realize logic functions by the    |                                                                         |
      |        | composition of simpler components. It can be seen as a form of Boolean division,       |                                                                         |
      |        | where a given function is split into a divisor and quotient (and a remainder, if needed). |                                                                         |
      |        | The key questions are how to find a good divisor and then how to compute the quotient. |                                                                         |
      |        | In this paper we choose as divisor an approximation of the given function, and          |                                                                         |
      |        | characterize the incompletely specified function which describes the full flexibility   |                                                                         |
      |        | for the quotient. We report at the end preliminary experiments for bi-decomposition      |                                                                         |
      |        | based on two AND-like operators with a divisor approximation from 1 to 6, and           |                                                                         |
      |        | discuss the impact of the approximation error rate on the final area of the components  |                                                                         |
      |        | in synthesis by three-level XOR-AND-OR forms.                                         |                                                                         |
      |        | Download Paper (PDF; Only available from the DATE venue WiFi)                        |                                                                         |
09:45  | 5.6.4  | MINIDELAY: MULTI-STRATEGY TIMING-AWARE LAYER ASSIGNMENT FOR ADVANCED TECHNOLOGY NODES  | Xinghai Zhang, Fuzhou University, CN                                    |
      |        | Speaker: Xinghai Zhang¹, Zhen Zhuang¹, Genggeng Liu¹, Xing Huang², Wen-Hao Liu¹,       |                                                                        |
      |        | Wenzhong Guo¹ and Ting-Chi Wang²                                                      |                                                                        |
      |        | ¹Fuzhou University, CN; ²National Tsing Hua University, TW; ³Cadence Design Systems, US |                                                                        |
      |        | Abstract                                                                             |                                                                         |
      |        | Layer assignment, a major step in global routing of integrated circuits, is usually     |                                                                         |
      |        | performed to assign segments of nets to multiple layers. Besides the                   |                                                                         |
      |        | traditional optimization goals such as overflow and via count, interconnect delay plays |                                                                         |
      |        | an important role in determining chip performance and has been                        |                                                                         |
      |        | attracting much attention in recent years. Accordingly, in this paper, we propose     |                                                                         |
      |        | MiniDelay, a timing-aware layer assignment algorithm to minimize delay for               |                                                                         |
      |        | advanced technology nodes, taking both wire congestion and coupling effect into        |                                                                         |
      |        | account. MiniDelay consists of the following three key techniques: 1) a non-            |                                                                         |
      |        | default-rule routing technique is adopted to reduce the delay of timing critical nets,  |                                                                         |
      |        | 2) an effective congestion assessment method is proposed to optimize delay of           |                                                                         |
      |        | nets and via count simultaneously, and 3) a net scalpel technique is proposed to        |                                                                         |
      |        | further reduce the maximum delay of nets, so that the chip                              |                                                                         |
      |        | performance can be improved in a global manner. Experimental results on                 |                                                                         |
      |        | multiple benchmarks confirm that the proposed algorithm leads to lower delay            |                                                                         |
      |        | and few vias, while achieving the best solution quality among the existing              |                                                                         |
      |        | algorithms with the shortest runtime.                                                |                                                                         |
      |        | Download Paper (PDF; Only available from the DATE venue WiFi)                        |                                                                         |
10:00  | IP2-    | A SCALABLE MIXED SYNTHESIS FRAMEWORK FOR HETEROGENEOUS NETWORKS                      | Max Austin, University of Utah, US                                      |
      | 16, 932 | Speaker: Max Austin¹, Scott Temple¹, Walter Lau Neto¹, Luca Amaru¹, Xifan Tang¹ and   |                                                                        |
      |        | Pierre-Emmanuel Gaillardon¹                                                            |                                                                        |
      |        | ¹University of Utah, US; ²Synopsys, US                                               |                                                                        |
      |        | Abstract                                                                             |                                                                         |
      |        | We present a new logic synthesis framework which produces efficient post-technology    |                                                                         |
      |        | mapped results on heterogeneous networks containing a mix of different types of logic.  |                                                                         |
      |        | This framework accomplishes this by breaking down the circuit into sections using a     |                                                                         |
      |        | hypergraph k-way partitioner and then determines the best-fit logic representation for  |                                                                         |
      |        | each partition between two Boolean networks, And-Inverter Graphs(AIG) and Majority-Inverter Graphs(MIG), |                                                                         |
      |        | which have been shown to perform better over each other on different types of logic.   |                                                                         |
      |        | Experimental results show that over a set of Open Piton DesignBenchmarks(OPDB) and     |                                                                         |
      |        | OpenCores benchmarks, our proposed methodology outperforms state-of-the-art academic   |                                                                         |
      |        | tools inArea-Delay Product(ADP), Power-Delay Product(PDP), and Energy-Delay Product(EDP)|                                                                         |
      |        | by 5%, 2%, and 15% respectively after performing Application Specific Integrated       |                                                                         |
      |        | Circuits(ASIC) technology mapping as well as showing a 54% improvement in runtime over  |                                                                         |
      |        | conventional MIG optimization.                                                        |                                                                         |
      |        | Download Paper (PDF; Only available from the DATE venue WiFi)                        |                                                                         |
10:01  | IP2-    | DISCERN: DISTILLING STANDARD CELLS FOR EMERGING RECONFIGURABLE NANOTECHNOLOGIES       | Shubham Rai, TU Dresden, DE                                             |
      | 16, 456 | Speaker: Shubham Rai¹, Michael Ratzza², Siva Satyendra Sahoo¹ and Akash Kumar¹        |                                                                        |
      |        | ¹TU Dresden, DE; ²TU Dresden and CfAED, DE                                           |                                                                        |
      |        | Abstract                                                                             |                                                                         |
      |        | Logic gates and circuits based on reconfigurable nanotechnologies demonstrate           |                                                                         |
      |        | runtime-reconfigurability, where a single logic gate can exhibit more than one         |                                                                         |
      |        | functionality. Recent attempts on circuits based on emerging reconfigurable              |                                                                         |
      |        | nanotechnologies have primarily focused on using the traditional CMOS design flow      |                                                                         |
      |        | involving similar-styled standard-cells. These CMOS-centric standard-cells fail to      |                                                                         |
      |        | utilize the exciting properties offered by these nanotechnologies. In the present      |                                                                         |
      |        | work, we explore the boolean properties that define the reconfigurable properties of a  |                                                                         |
      |        | logic gate. By analyzing the truth-table in detail, we find that there is a common      |                                                                         |
      |        | boolean rule which dictates why a logic gate is reconfigurable. Such logic gates can    |                                                                         |
      |        | be efficiently implemented using reconfigurable nanotechnologies. We propose an        |                                                                         |
      |        | algorithm which analyses the truth-tables of nodes in a circuit to list all potential  |                                                                         |
      |        | reconfigurable logic gates for a particular circuit. Technology mapping with these new |                                                                         |
      |        | logic gates (or standard-cells) leads to a better mapping in terms of area and delay.   |                                                                         |
      |        | Experiments employing our methodology over E9PL benchmarks, show average improvements  |                                                                         |
      |        | of around 13% and 15% in terms of area, number of edges and delay respectively as       |                                                                         |
      |        | compared to the conventional CMOS-centric standard-cell based mapping.                |                                                                         |
      |        | Download Paper (PDF; Only available from the DATE venue WiFi)                        |                                                                         |
10:00  |        | End of session                                                                       |                                                                         |

5.7 Stochastic Computing

Date: Wednesday 11 March 2020
Time: 08:30 - 10:00
Location / Room: Berlioz
Chair: Robert Wille, Johannes Kepler University Linz, AT
Co-Chair: Shigeru Yamashita, Ritsumeikan, JP

Stochastic computing uses random bitstreams to reduce computational and area costs of a general class of Boolean operations, including arithmetic addition and multiplication. This session considers stochastic computing from a model-, accuracy-, and application-perspective, by presenting papers that span from models of pseudo-random number generators, to accuracy analysis of stochastic circuits, to novel applications for signal processing tasks.
10:00  5.7.2 ACCURACY ANALYSIS FOR STOCHASTIC CIRCUITS WITH D-FLIP FLOP INSERTION
Speaker: Kuncai Zhong, University of Michigan-Shanghai Jiao Tong University Joint Institute, CN
Authors: Kuncai Zhong and Weikang Qian, Shanghai Jiao Tong University, CN
Abstract
One of the challenges stochastic computing (SC) faces is the high cost of stochastic number generators (SNG). A solution to it is inserting D flip-flops (DFFs) into the circuit. However, the accuracy of the stochastic circuits would be affected and it is crucial to capture it. In this work, we propose an efficient method to analyze the accuracy of stochastic circuits with DFFs inserted. Furthermore, given the importance of multiplication, we apply this method to analyze stochastic multiplier with DFFs inserted. Several interesting claims are obtained about the use of probability conversion circuits. For example, using weighted binary generator is more accurate than using comparator. The experimental results show the correctness of the proposed method and the claims. Furthermore, the proposed method is up to 560× faster than the simulation-based method.
Download Paper (PDF; Only available from the DATE venue WiFi)

09:30  5.7.3 DYNAMIC STOCHASTIC COMPUTING FOR DIGITAL SIGNAL PROCESSING APPLICATIONS
Speaker: Jie Han, University of Alberta, CA
Authors: Siting Liu and Jie Han, University of Alberta, CA
Abstract
Stochastic computing (SC) utilizes a random binary bit stream to encode a number by counting the frequency of 1’s in the stream (or sequence). Typically, a small circuit is used to perform a bit-wise logic operation on the stochastic sequences, which leads to significant hardware and power savings. Energy efficiency, however, is a challenge for SC due to the long sequences required for accurately encoding numbers. To overcome this challenge, we consider to use a stochastic sequence to encode a continuously variable signal instead of a number to achieve higher accuracy, higher energy efficiency and greater flexibility. Specifically, one single bit is used to encode a sample from a signal for efficient processing. This type of sequences encodes constantly variable values, so it is referred to as dynamic stochastic sequences (DSS’s). The DSS enables the use of SC circuits to efficiently perform tasks such as frequency mixing and function estimation. It is shown that such a dynamic SC (DSC) system achieves savings up to 98.4% in energy and up to 96.8% in time with a slightly higher accuracy compared to conventional SC. It also achieves energy and time savings of up to 60% compared to a fixed-width binary implementation.
Download Paper (PDF; Only available from the DATE venue WiFi)

10:00  5.1.4 A 16×128 STOCHASTIC-BINARY PROCESSING ELEMENT ARRAY FOR ACCELERATING STOCHASTIC DOT-PRODUCT COMPUTATION USING 1-16 BIT-STREAM LENGTH
Speaker: Hyunjoon Kim, Nanyang Technological University, SG
Authors: Qian Chen, Yuqi Su, Hyunjoon Kim, Taegeun Yoo, Tony Tae-Hyoung Kim and Bongjin Kim, Nanyang Technological University, SG
Abstract
This work presents 16×128 stochastic-binary processing element array for accelerating stochastic dot-product computation. The processing element (PE) with all-digital components consists of an XOR gate as a bipolar stochastic multiplier and an 8bit binary adder with 8× registers for accumulating partial-sums. The PE array comprises 16× dot-product units, each with 128 PEs cascaded in a single row. The latency and energy of the proposed dot-product unit is minimized by reducing the number of bit-streams required for minimizing the accuracy degradation induced by the approximate stochastic computing. A 128-input dot-product operation requires the bit-stream length (N) of 1-to-16, which is two orders of magnitude smaller than the baseline stochastic computation using MUX-based adders. The simulated dot-product error is 6.9-to-1.5% for N=1-to-16, while the error from the baseline stochastic method is 5.9-to-1.7% with N=128-to-2048. A mean MNIST classification accuracy is 96.11% (which is 1.19% lower than 8b binary) using a three-layer MLP at N=16. The measured energy from a 65nm test-chip is 10.04pJ per dot-product, and the energy efficiency is 25.5TOPS/W at N=16.
Download Paper (PDF; Only available from the DATE venue WiFi)

10:01  5.1.6 TOWARDS EXPLORING THE POTENTIAL OF ALTERNATIVE QUANTUM COMPUTING ARCHITECTURES
Speaker: Arighna Deb, Kalinga Institute of Industrial Technology, IN
Authors: Arighna Deb1, Gerhard W. Dueck2 and Robert Wille3
1Kalinga Institute of Industrial Technology, IN; 2University of New Brunswick, CA; 3Hannes Kepler University Linz, AT
Abstract
The recent advances in the physical realization of Noisy Intermediate Scale Quantum (NISQ) computers have motivated research on design automation that allows users to execute quantum algorithms on them. Certain physical constraints in the architectures restrict how logical qubits used to describe the algorithm can be mapped to physical qubits used to realize the corresponding functionality. Thus far, this has been addressed by inserting additional operations in order to overcome the physical constraints. However, all these approaches have taken the existing architectures as invariant and did not explore the potential of changing the quantum architecture itself—a valid option as long as the underlying physical constraints remain satisfied. In this work, we propose initial ideas to explore this potential. More precisely, we introduce several schemes for the generation of alternative coupling graphs (and, by this, quantum computing architectures) that still might be able to satisfy physical constraints but, at the same time, allow for a more efficient realization of the desired quantum functionality.
Download Paper (PDF; Only available from the DATE venue WiFi)
5.7 Special Session: HLS for AI HW

Date: Wednesday 11 March 2020
Time: 08:30 - 10:00
Location / Room: Exhibition Theatre

Chair:
Massimo Cecchetti, Mentor, A Siemens Business, US

Co-Chair:
Astrid Ernst, Mentor, A Siemens Business, US

One of the fastest growing areas of hardware and software design is artificial intelligence (AI)/machine learning (ML), fueled by the demand for more autonomous systems like self-driving vehicles and voice recognition for personal assistants. Many of these algorithms rely on convolutional neural networks (CNNs) to implement deep learning systems. While the concept of convolution is relatively straightforward, the application of CNNs to the ML domain has yielded dozens of different neural network approaches. These networks can be executed in software on CPUs/GPUs, the power requirements for these solutions make them impractical for most inferencing applications, the majority of which involve portable, low-power devices. To improve the power/performance, hardware teams are forming to create ML hardware acceleration blocks. However, the process of taking any one of these compute-intensive networks into hardware, especially energy-efficient hardware, is a time consuming process if the team employs a traditional RTL design flow. Consider all of these interdependent activities using a traditional flow: • Expressing the algorithm correctly in RTL. • Choosing the optimal bit-widths for kernel weights and local storage to meet the memory budget. • Designing the microarchitecture to have a low enough latency to be practical for the target application, while determining how the interconnect communicates across the system bus without killing the latency the team just fought for. • Verifying the algorithm early on and throughout the implementation process, especially in the context of the entire system. • Optimizing for power for mobile devices. • Getting the product to market on time. This domain is in desperate need of a productivity-booster methodology shift away from an RTL flow.
IP2-1

**SAMPLING FROM DISCRETE DISTRIBUTIONS IN COMBINATIONAL HARDWARE WITH APPLICATION TO POST-QUANTUM CRYPTOGRAPHY**

**Speaker:** Michael Lyons, George Mason University, US

**Authors:** Michael Lyons and Kris Gej, George Mason University, US

**Abstract**

Random values from discrete distributions are typically generated from uniformly-random samples. A common technique is to use a cumulative distribution table (CDT) lookup for inversion sampling, but it is also possible to use Boolean functions to map a uniformly-random bit sequence into a value from a discrete distribution. This work presents a methodology for deriving such functions for any discrete distribution, encoding them in VHDL for implementation in combinational hardware, and (for moderate precision and sample space size) confirming the correctness of the produced distribution. The process is demonstrated using a discrete Gaussian distribution with a small sample space, but it is applicable to any discrete distribution with fixed parameters. Results are presented for sampling schemes from several submissions to the NIST PQC standardization process, confirming this methodology against a DCT lookups on a Xilinx Artix-7 FPGA. The process produces compact solutions for distributions up to moderate size and precision.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

IP2-2

**ON THE PERFORMANCE OF NON-PROFILED DIFFERENTIAL DEEP LEARNING ATTACKS AGAINST AN AES CRYPTOGRAPHY ALGORITHM PROTECTED USING A CORRELATED NOISE HIDING COUNTERMEASURE**

**Speaker:** Amir Alipour, Grenoble INP ESIAS, FR

**Authors:** Amir Alipour, Athanasios Papadimitriou, Vincent Berouille, Ehsan Aerabi and David Hely

1University Grenoble Alpes, Grenoble INP ESISAR, LCS Laboratory, FR; 2University Grenoble Alpes, Grenoble INP ESISAR, ESYNOV, FR; 3University Grenoble Alpes, Grenoble INP ESISAR, LSC Laboratory, FR

**Abstract**

Recent works in the field of cryptography focus on Deep Learning based Side Channel Analysis (DLSCA) as one of the most powerful attacks against common encryption algorithms such as AES. As a common case, profiling DLSCA have shown great capabilities in revealing secret cryptographic keys against the majority of AES implementations. In a very recent study, it has been shown that Deep Learning can be applied in a non-profiling way (non-profiling DLSCA), making this method considerably more practical, and able to break powerful countermeasures for encryption algorithms such as AES including masking countermeasures, requiring considerably less power traces than a first order CPA attack. In this work, our main goal is to apply the non-profiling DLSCA against a hiding-based AES countermeasure which utilizes correlated noise generation so as to hide the secret encryption key. We show that this AES, with correlated noise generation as a lightweight countermeasure, can provide equivalent protection under CPA and under non-profiling DLSCA attacks, in terms of the required power traces to obtain the secret key.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

IP2-3

**FAST AND ACCURATE PERFORMANCE EVALUATION FOR RISC-V USING VIRTUAL PROTOTYPES**

**Speaker:** Vladimir Herdt, University of Bremen, DE

**Authors:** Vladimir Herdt, Daniel Grosse and Rolf Drechsler

1University of Bremen, DE; 2University of Bremen / DFKI, DE

**Abstract**

RISC-V is gaining huge popularity in particular for embedded systems. Recently, a SystemC-based Virtual Prototype (VP) has been open sourced to lay the foundation for providing support for system-level use cases such as design space exploration, analysis of complex HW/SW interactions and power/timing/performance validation for RISC-V based systems. In this paper, we propose an efficient core timing model and integrate it into the VP core to enable fast and accurate performance evaluation for RISC-V based systems. As a case-study we provide a timing configuration matching the RISC-V HiFive1 board from SiFive. Our experiments demonstrate that our approach allows to obtain very accurate performance evaluation results while still retaining a high simulation performance.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

IP2-4

**AUTOMATED GENERATION OF LTL SPECIFICATIONS FOR SMART HOME IOT USING NATURAL LANGUAGE**

**Speaker:** Shiyu Zhang, Nanjing University, CN

**Authors:** Shiyu Zhang, Juan Zhai, Lei Bu, Mingsong Chen, Linzhang Wang and Xuandong Li

1Nanjing University, CN; 2East China Normal University, CN

**Abstract**

Ordinary inexperienced users can build their smart home IoT system easily nowadays, but such user-customized systems could be error-prone. Using formal verification to prove the correctness of such systems is necessary. However, to conduct formal proof, formal specifications such as Linear Temporal Logic (LTL) formulas have to be provided, but ordinary users cannot author-LTL formulas but only natural language. To address this problem, this paper presents a novel approach that can automatically generate formal LTL specifications from natural language requirements based on domain knowledge and our proposed ambiguity refining techniques. Experimental results show that our approach can achieve a high correctness rate of 95.4% in converting natural language sentences into LTL formulas from 481 requirements of real examples.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

IP2-5

**A HEAT-RECIRCULATION-AWARE VM PLACEMENT STRATEGY FOR DATA CENTERS**

**Speaker:** Hao Feng, Yuhei Deng and Yi Zhou

1Jinan University, CN; 2Chinese Academy of Sciences; Jinan University, CN; 3Columbus State University, US

**Abstract**

Data centers consisted of a great number of IT devices (e.g., servers, switches and etc.) which generates a massive amount of heat emission. Due to the special arrangement of racks in the data center, heat recirculation often occurs between nodes. It can cause a sharp rise in temperature of the equipment coupled with local hot spots in data centers. Existing VM placement strategies can minimize energy consumption of data centers by optimizing resource allocation in terms of multiple physical resources (e.g., memory, bandwidth, cpu and etc.). However, existing strategies ignore the role of heat recirculation in the data center. To address this problem, in this study, we propose a heat-recirculation-aware VM placement strategy and design a Simulated Annealing Based Algorithm (SABA) to lower the energy consumption of data centers. Different from the existing SA algorithm, SABA optimize the distribution of the initial solution and the way of iteration. We quantitatively evaluate SABA’s performance in terms of algorithm efficiency, the activated servers and the energy saving against with XINT-GA algorithm (Thermal-aware task scheduling strategy), FCFS (First-Come First-Served) and SA. Experimental results indicate that our heat-recirculation-aware VM placement strategy provides a powerful solution for improving energy efficiency of data centers.

**Download Paper (PDF; Only available from the DATE venue WiFi)**
**IP2-6 ENERGY OPTIMIZATION IN NCFET-BASED PROCESSORS**

**Authors:** Sami Salamini1, Martin Rapp1, Hussam Amrouch1, Andreas Gerstlauer1 and Joerg Henkel1

1Karlsruhe Institute of Technology, DE; 2University of Texas at Austin, US

**Abstract**

Energy consumption is a key optimization goal for all modern processors. Negative Capacitance Field-Effect Transistors (NCFETs) are a leading emerging technology that promises outstanding performance in addition to better energy efficiency. The thickness of the additional ferroelectric layer, frequency, and voltage are the key parameters in NCFET technology that impact the power and frequency of processors. However, their joint impact on energy optimization has not been investigated yet. In this work, we are the first to demonstrate that conventional (i.e., NCFET-unaware) dynamic voltage/frequency scaling (DVFS) techniques to minimize energy are sub-optimal when applied to NCFET-based processors. We further demonstrate that state-of-the-art NCFET-aware voltage scaling for power minimization is also sub-optimal when it comes to energy. This work provides the first NCFET-aware DVFS technique that optimizes the processor’s energy through optimal runtime frequency/voltage selection. In NCFETs, energy-optimal frequency and voltage are dependent on the workload and technology parameters. Our NCFET-aware DVFS technique considers these effects to perform optimal voltage/frequency selection at runtime depending on workload characteristics. Results show up to 90 % energy savings compared to conventional DVFS techniques. Compared to state-of-the-art NCFET-aware power management, our technique provides up to 72 % energy savings along with 3:7x higher performance.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**IP2-7 TOWARDS A MODEL-BASED MULTI-OBJECTIVE OPTIMIZATION APPROACH FOR SAFETY-CRITICAL REAL-TIME SYSTEMS**

**Speaker:** Emmanuel Grolleau, LIAS / ISAE-ENMSA, FR

**Authors:** Soulimane Kamni1, Yassine OUHAMMOU2, Antoine Bertout3 and Emmanuel Grolleau4

1LIAS/ENMSA, FR; 2LIAS / ISAE-ENMSA, FR; 3LIAS, Université de Poitiers, ISAE-ENMSA, FR; 4LIAS, ISAE-ENMSA, Université de Poitiers, FR

**Abstract**

In safety-critical real-time systems domain, obtaining the appropriate operational model which meets the temporal (e.g. deadlines) and business (e.g. redundancy) requirements while being optimal in terms of several metrics is a primalordial process in the design life-cycle. Recently, several researches have proposed to explore cross-domain trade-offs for a higher behaviour performance. Indeed, this process represents the first step in the deployment phase, which is very sensitive because it could be error-prone and time consuming. This paper is a work in progress proposing an approach aiming to help real-time system architects to take benefit from existing works, overcome their limits, and capitalize the efforts. Furthermore, the approach is based on the model-driven engineering paradigm and suggests to ease the usage of methods and tools thanks to repositories gathering them as a sort of a shared knowledge.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**IP2-8 CURRENT-MODE CARRY-FREE MULTIPLIER DESIGN USING A MEMRISTOR-TRANSISTOR CROSSBAR ARCHITECTURE**

**Speaker:** Shengji1 Yu, Newcastle University, GB

**Authors:** Shengji1 Yu, Ahmed Soltan2, Rishad Shafi2, Thanasis Bunn11, Domenico Balsamo1, Fei Xia1 and Alex Yakovlev1

1Newcastle University, GB; 2Nile University, EG

**Abstract**

Traditional multipliers consist of complex logic components. They are a major energy and performance contributor of modern computer-intensive applications. As such, designing multipliers with reduced energy and faster speed has remained a thoroughgoing challenge. This paper presents a novel, carry-free multiplier, which is suitable for new-generation of energy-constrained applications. The multiplier circuit consists of an array of memristor-transistor cells that can be selected (i.e., turned ON or OFF) using a combination of DC bias voltages based on the operand values. When a cell is selected it contributes to current in the array path, which is then amplified by current mirrors with variable transistor gate sizes. The different current paths are connected to a node for analogously accumulating the currents to produce the multiplier output directly, removing the carry propagation stages, typically seen in traditional digital multipliers. An essential feature of this multiplier is autonomous survivability, i.e., when the power is below this threshold the logic state automatically retains at a zero-cost due to the non-volatile properties of memristors.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**IP2-9 N-BIT DATA PARALLEL SPIN WAVE LOGIC GATE**

**Speaker:** Abdulqader Mahmoud1, TU Delft, NL

**Authors:** Abdulqader Mahmoud1, Frederic Vanderveken1, Florin Ciubotaru1, Christoph Adelmann1, Sorin Cotofana1 and Said Hamedi1

1TU Delft, NL; 2IMEC, BE

**Abstract**

Due to their very nature, Spin Waves (SWs) created in the same waveguide, but with different frequencies, can coexist while selectively interacting with their own species only. The absence of inter-frequency interferences isolates input data sets encoded in SWs with different frequencies and creates the premises for simultaneous data parallel SW based processing without hardware replication or delay overhead. In this paper we leverage this SW property by introducing a novel computation paradigm, which allows for the parallel processing of n-bit input data vectors on the same basic SW based logic gate. Subsequently, to demonstrate the proposed concept, we present 8-bit parallel 3-input Majority gate implementation and validate it by means of Object Oriented Micromagnetic Framework (OOMMF) simulations. To evaluate the potential benefit of our proposal we compare the 8-bit data parallel gate with equivalent scalar SW gate based implementation. Our evaluation indicates that 8-bit data 3-input Majority gate implementation requires 4.16x less area than the scalar SW gate based equivalent counterpart while preserving the same delay and energy consumption figures.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**IP2-10 HIGH-SPEED ANALOG SIMULATION OF CMOS VISION CHIPS USING EXPLICIT INTEGRATION TECHNIQUES ON MANY-CORE PROCESSORS**

**Speaker:** Tom Kazmierski, University of Southampton, GB

**Authors:** Gines Domenech-Arns1 and Tom J Kazmierski2

1Universidad Politecnica de Cartagena, ES; 2University of Southampton, GB

**Abstract**

This work describes a high-speed simulation technique of analog circuits which is based on the use of state-space equations and an explicit integration method parallelised on a multiprocessor architecture. The integration step of such method is smaller than the one required by an implicit simulation technique based on Newton-Raphson iterations. However, given that explicit methods do not require the computation of time-consuming matrix factorizations, the overall simulation time is reduced. The technique described in this work has been implemented on a NVIDIA general purpose GPU and has been tested simulating the Gaussian filtering operation performed by a smart CMOS image sensor. Such devices are used to perform computation on the edge and include built-in image processing functions. Among those, the Gaussian filtering is one of the most common functions, since it is a basic task for early vision processing. These smart sensors are increasingly complex and hence the time required to simulate them during their design cycle is also larger and larger. From a certain imager size, the proposed simulation method yields simulation times two order of magnitude faster than an implicit method based tool such as SPCI.

**Download Paper (PDF; Only available from the DATE venue WiFi)**
A 100KHz-1GHz TERMINATION-DEPENDENT HUMAN BODY COMMUNICATION CHANNEL MEASUREMENT USING MINIATURIZED WEARABLE DEVICES

**Speaker:** Shreyas Sen, Purdue University, US

**Authors:** Shitij Avlani, Mayukh Nath, Shovan Maity and Shreyas Sen, Purdue University, US

**Abstract**

Human Body Communication has shown great promise to replace wireless communication for information exchange between wearable devices of a body area network. However, there are very few studies in literature, that systematically study the channel loss of capacitive HBC for wearable devices over a wide frequency range with different terminations at the receiver, partly due to the need for miniaturized wearable devices for an accurate study. This paper, for the first time, measures the channel loss of capacitive HBC from 100KHz to 1GHz for both high-impedance and 50 ohm terminations using wearable, battery powered devices; which is mandatory for accurate measurement of the HBC channel-loss, due to ground coupling effects. Results show that high impedance termination leads to a significantly lower channel loss (40 dB improvement at 1MHz), as compared to 50 ohm termination at low frequencies. This difference steadily decreases with increasing frequency, until they become similar near 80MHz. Beyond 100MHz inter-device coupling dominates, thereby preventing accurate measurements of channel loss of the human body. The measured results provide a consistent wearable, wide-frequency HBC channel loss data and could serve as a backbone for the emerging field of HBC by aiding in the selection of an appropriate operation frequency and termination.

Download Paper (PDF; Only available from the DATE venue WiFi)

FROM DRUP TO PAC AND BACK

**Speaker:** Daniela Kaufmann, Johannes Kepler University Linz, AT

**Authors:** Daniela Kaufmann, Armin Biere and Manuel Kauers, Johannes Kepler University Linz, AT

**Abstract**

Currently the most efficient automatic approach to verify gate-level multipliers combines SAT solving and computer algebra. In order to increase confidence in the verification, proof certificates are generated. However, due to different solving techniques, these certificates require two different proof formats, namely DRUP and PAC. A combined proof has so far been missing. Correctness of this approach can thus only be trusted up to the correctness of compositional reasoning. In this paper we show how to generate a single proof in one proof format, which then allows to certify correctness using one simple proof checker. We further investigate empirically the effect on proof generation and checking time as well as on proof size. It turns out that PAC proofs are much more compact and faster to check.

Download Paper (PDF; Only available from the DATE venue WiFi)

VERIFIABLE SECURITY TEMPLATES FOR HARDWARE

**Speaker:** Bill Harrison, Oak Ridge National Laboratory, US

**Authors:**

William Harrison1 and Gerard Alwein2

1Oak Ridge National Laboratory, US; 2Naval Research Laboratory, US

**Abstract**

But HLS has, with a few notable exceptions, not focused on transferring ideas and techniques from high assurance software formal methods to hardware development, despite there being a sophisticated and mature body of research in that area. Just as it has introduced software engineering virtues, we believe HLS can also become a vector for retrofitting software formal methods to the challenge of high assurance security in hardware. This paper introduces the Device Calculus and its mechanization in the Agda proof checking system. The Device Calculus is a starting point for exploring formal methods and security within high-level synthesis flows. We illustrate the Device Calculus with a number of examples of formally verifiable security templates---i.e., functions in the Device Calculus that express common security structures at a high-level of abstraction.

Download Paper (PDF; Only available from the DATE venue WiFi)

IFFSET: IN-FIELD FUZZING OF INDUSTRIAL CONTROL SYSTEMS USING SYSTEM EMULATION

**Speaker:** Dimitrios Tychalas, New York University, US

**Authors:** Dimitrios Tychalas1 and Michail Maniatakos2

1New York University, US; 2New York University Abu Dhabi, AE

**Abstract**

Industrial Control Systems (ICS) have evolved in the last decade, shifting from proprietary software/hardware to contemporary embedded architectures paired with open-source operating systems. In contrast to the IT world, where continuous updates and patches are expected, decommissioning always-on ICS for security assessment can incur prohibitive costs to their owner. Thus, a solution for routinely assessing the cybersecurity posture of diverse ICS without affecting their operation is essential. Therefore, in this paper we introduce IFFSET, a platform that leverages full system emulation of Linux-based ICS firmware and utilizes fuzzing for security evaluation. Our platform extracts the file system and kernel information from a live ICS device, building an image which is emulated on a desktop system through QEMU. We employ fuzzing as a security assessment tool to analyze ICS specific libraries and find potential security threatening conditions. We test our platform with commercial PLCs, showcasing potential threats with no interruption to the control process.

Download Paper (PDF; Only available from the DATE venue WiFi)

FANNET: FORMAL ANALYSIS OF NOISE TOLERANCE, TRAINING BIAS AND INPUT SENSITIVITY IN NEURAL NETWORKS

**Speaker:** Mahum Naseer, TU Wien, AT

**Authors:**

Mahum Naseer1, Mishal Fatima Minhas2, Faiq Khalid1, Muhammad Abdullah Hanif1, Osman Hasan2 and Muhammad Shafique1

1TU Wien, AT; 2National University of Sciences and Technology, PK

**Abstract**

With a constant improvement in the network architectures and training methodologies, Neural Networks (NNs) are increasingly being deployed in real-world Machine Learning systems. However, despite their impressive performance on "known inputs", these NNs can fail absurdly on the "unseen inputs", especially if these real-time inputs deviate from the training dataset distributions, or contain certain types of input noise. This indicates the low noise tolerance of NNs, which is a major reason for the recent increase of adversarial attacks. This is a serious concern, particularly for safety-critical applications, where inaccurate results lead to dire consequences. We propose a novel methodology that leverages model checking for the Formal Analysis of Neural Network (FANNet) under different input noise ranges. Our methodology allows us to rigorously analyze the noise tolerance of NNs, their input node sensitivity, and the effects of training bias on their performance, e.g., in terms of classification accuracy. For evaluation, we use a feed-forward fully-connected NN architecture trained for the Leukemia classification. Our experimental results show 11% noise tolerance for the given trained network, identify the most sensitive input nodes, confirm the biasness of the available training dataset, and indicate that the proposed methodology is much more rigorous and yet comparable to validation testing in terms of time and computational resources for larger noise ranges.

Download Paper (PDF; Only available from the DATE venue WiFi)
A SCALABLE MIXED SYNTHESIS FRAMEWORK FOR HETEROGENEOUS NETWORKS

**Speaker:** Max Austin, University of Utah, US

**Authors:**
Max Austin1, Scott Temple1, Walter Lau Neto1, Luca Amaru2, Xifan Tang1 and Pierre-Emmanuel Gaillardon1

1University of Utah, US; 2Synopsys, US

**Abstract**
We present a new logic synthesis framework which produces efficient post-technology mapped results on heterogeneous networks containing a mix of different types of logic. This framework accomplishes this by breaking down the circuit into sections using a hypergraph k-way partitioner and then determines the best-fit logic representation for each partition between two Boolean networks, And-Inverter Graphs (AIG) and Majority-Inverter Graphs (MIG), which have been shown to perform better over each other on different types of logic. Experimental results show that over a set of Open Piton DesignBenchmarks (OPDB) and OpenCores benchmarks, our proposed methodology outperforms state-of-the-art academic tools in Area-Delay Product (ADP), Power-Delay Product (PDP), and Energy-Delay Product (EDP) by 5%, 2%, and 15% respectively after performing Application Specific Integrated Circuits (ASIC) technology mapping as well as showing a 54% improvement in runtime over conventional MIG optimization

Download Paper (PDF; Only available from the DATE venue WiFi)

DISCERN: DISTILLING STANDARD CELLS FOR EMERGING RECONFIGURABLE NANOTECHNOLOGIES

**Speaker:** Shubham Rai, TU Dresden, DE

**Authors:**
Shubham Rai1, Michael Raizt2, Siva Satyendra Sahoo1 and Akash Kumar1

1TU Dresden, DE; 2TU Dresden and CAED, DE

**Abstract**
Logic gates and circuits based on reconfigurable nanotechnologies demonstrate runtime-reconfigurability, where a single logic gate can exhibit more than one functionality. Recent attempts on circuits based on emerging reconfigurable nanotechnologies have primarily focused on using the traditional CMOS design flow involving similar-standard-cells. These CMOS-centric standard-cells fail to utilize the exciting properties offered by these nanotechnologies. In the present work, we explore the boolean properties that define the reconfigurable properties of a logic gate. By analyzing the truth-table in detail, we find that there is a common boolean rule which dictates why a logic gate is reconfigurable. Such logic gates can be efficiently implemented using reconfigurable nanotechnologies. We propose an algorithm which analyses the truth-tables of nodes in a circuit to list all such potential reconfigurable logic gates for a particular circuit. Technology mapping with these new logic gates (or standard-cells) leads to a better mapping in terms of area and delay. Experiments employing our methodology over EPFL benchmarks, show average improvements of around 13%, 16% and 11.5% in terms of area, number of edges and delay respectively as compared to the conventional CMOS-centric standard-cell based mapping.

Download Paper (PDF; Only available from the DATE venue WiFi)

A 16x128 STOCHASTIC-BINARY PROCESSING ELEMENT ARRAY FOR ACCELERATING STOCHASTIC DOT-PRODUCT COMPUTATION USING 1-16 BIT-STREAM LENGTH

**Speaker:** Hyunjoon Kim, Nanyang Technological University, SG

**Authors:**
Qian Chen, Yuqi Su, Hyunjoon Kim, Taegeun Yoo, Tony Tae-Hyung Kim and Bongjin Kim, Nanyang Technological University, SG

**Abstract**
This work presents 16x128 stochastic-binary processing elements for energy/area efficient processing of artificial neural networks. A processing element (PE) with all-digital components consists of an XNOR gate as a bipolar stochastic multiplier and an 8bit binary adder with 8x registers for accumulating partial-sums. The PE array comprises 16x dot-product units, each with 128 PEs cascaded in a single row. The latency and energy of the proposed dot-product unit is minimized by reducing the number of bit-streams required for minimizing the accuracy degradation induced by the approximate stochastic computing. A 128-input dot-product operation requires the bit-stream length (N) of 1-to-16, which is two orders of magnitude smaller than the baseline stochastic computation using MUX-based adders. The simulated dot-product error is 6.9-to-1.5% for N=1-to-16, while the error from the baseline stochastic method is 5.9-to-1.7% with N=128-to-2048. A mean MNIST classification accuracy is 96.11% (which is 1.19% lower than 8 bit binary) using a three-layer MLP at N=16. The measured energy from a 65nm test-chip is 10.04pJ per dot-product, and the energy efficiency is 25 STOPS/W at N=16.

Download Paper (PDF; Only available from the DATE venue WiFi)

TOWARDS EXPLORING THE POTENTIAL OF ALTERNATIVE QUANTUM COMPUTING ARCHITECTURES

**Speaker:** Arighna Deb, Kalinga Institute of Industrial Technology, IN

**Authors:**
Arighna Deb1, Gerhard W. Dueck2 and Robert Wille3

1Kalinga Institute of Industrial Technology, IN; 2University of New Brunswick, CA; 3Johannes Kepler University Linz, AT

**Abstract**
The recent advances in the physical realization of Noisy Intermediate Scale Quantum (NISQ) computers have motivated research on design automation that allows users to execute quantum algorithms on them. Certain physical constraints in the architectures restrict how logical gates used to describe the algorithm can be mapped to physical qubits used to realize the corresponding functionality. Thus far, this has been addressed by inserting additional operations in order to overcome the physical constraints. However, all these approaches have taken the existing architectures as invariant and did not explore the potential of changing the quantum architecture itself—a valid option as long as the underlying physical constraints remain satisfied. In this work, we propose initial ideas to explore this potential. More precisely, we introduce several schemes for the generation of alternative coupling graphs (and, by this, quantum computing architectures) that still might be able to satisfy physical constraints but, at the same time, allow for a more efficient realization of the desired quantum functionality.

Download Paper (PDF; Only available from the DATE venue WiFi)

ACCELERATING QUANTUM APPROXIMATE OPTIMIZATION ALGORITHM USING MACHINE LEARNING

**Speaker:** Swaroop Ghosh, Pennsylvania State University, US

**Authors:**
Mahabubul Alam, Abdullah Ash- Saki and Swaroop Ghosh, Pennsylvania State University, US

**Abstract**
We propose a machine learning based approach to accelerate quantum approximate optimization algorithm (QAOA) implementation which is a promising quantum-classical hybrid algorithm to prove the so-called quantum supremacy. In QAOA, a parametric quantum circuit and a classical optimizer iterate in a closed loop to solve hard combinatorial optimization problems. The performance of QAOA improves with an increasing number of stages (depth) in the quantum circuit. However, two new parameters are introduced with each added stage for the classical optimizer increasing the number of optimization loop iterations. We note a correlation among parameters of the lower-depth and the higher-depth QAOA implementations and, exploit it by developing a machine learning model to predict the gate parameters close to the optimal values. As a result, the optimization loop converges in a fewer number of iterations. We choose graph MaxCut problem as a prototype to solve using QAOA. We perform a feature extraction routine using 100 different QAOA instances and develop a training data-set with 13,860 optimal parameters. We present our analysis for 4 flavors of regression models and 4 flavors of classical optimizers. Finally, we show that the proposed approach can curtail the number of optimization iterations by on average 44.9% (up to 65.7%) from an analysis performed with 264 flavors of graphs.

Download Paper (PDF; Only available from the DATE venue WiFi)
**ABSTRACTION TO HYBRID AUTOMATA**

**Authors:**
Ahmad Tarraf and Lars Hedrich, University of Frankfurt, DE

Abstract

Model abstraction of transistor-level circuits, while preserving an accurate behavior, is still an open problem. In this demo an approach is presented that automatically generates a hybrid automaton (Ha) with linear states from an existing circuit netlist. The approach starts with a netlist at transistor level with full SPICE accuracy and ends at the system level description of the circuit in matlab or in Verilog-A. The resulting hybrid automaton exhibits linear behavior as well as the technology dependent nonlinear e.g. limiting behavior. The accuracy and speed-up of the Verilog-A generated models is evaluated based on several transistor level circuit abstractions of simple operational amplifiers up to a complex filters. Moreover, we verify the equivalence between the generated model and the original circuit. For the generated models in matlab syntax, a reachability analysis is performed using the reachability tool cora.

More information ...

**EUCLID-NIR GPU: AN ON-BOARD PROCESSING GPU-ACCELERATED SPACE CASE STUDY DEMONSTRATOR**

**Authors:**
Ivan Rodríguez and Leonidas Kosmidis, BSC / UPC, ES

Abstract

Embedded Graphics Processing Units (GPUs) are very attractive candidates for on-board payload processing of future space systems, thanks to their high performance and low-power consumption. Although there is significant interest from both academia and industry, there is no open and publicly available case study showing their capabilities yet. In this master thesis project, which was performed within the GPU4S (GPU for Space) ESA-funded project, we have parallelised and ported the Euclid NIR (Near Infrared) image processing algorithm used in the European Space Agency’s (ESA) mission to be launched in 2022, to an automotive GPU platform, the NVIDIA Xavier. In the demo we will present in real-time its significantly higher performance achieved compared to the original sequential implementation. In addition, visitors will have the opportunity to examine the images on which the algorithm operates, as well as to inspect the algorithm parallelisation through profiling and code inspection.

More information ...

**BCFLEAM: BACKFLOW: BACKWARD EDGE CONTROL FLOW ENFORCEMENT FOR LOW END ARM REAL-TIME SYSTEMS**

**Authors:**
Bresch Cyril¹, David Héy¹, Roman Lysecky² and Stephanie Chollet¹

¹LCIS, FR; ²University of Arizona, US

Abstract

The C programming language is one of the most popular languages in embedded system programming. Indeed, C is efficient, lightweight and can easily meet high performance and deterministic real-time constraints. However, these assets come at a certain price. Indeed, C does not provide extra features for memory safety. As a result, attackers can easily exploit spatial memory vulnerabilities to hijack the execution flow of an application. The demonstration features a real-time connected event-driven circuit abstraction of simple operational amplifiers up to a complex filters. Moreover, we verify the equivalence between the generated model and the original circuit. For the generated models in matlab syntax, a reachability analysis is performed using the reachability tool cora.

More information ...

**UWB ACKATCK: HIJACKING DEVICES IN UWB INDOOR POSITIONING SYSTEMS**

**Authors:**
Baptiste Pestourie, Vincent Berouille and Nicolas Fourty, Université Grenoble Alpes, FR

Abstract

Various radio-based Indoor Positioning Systems (IPS) have been proposed during the last decade as solutions to GPS inconsistency in indoor environments. Among the different radio technologies proposed for this purpose, 802.15.4 Ultra-Wideband (UWB) is by far the most performant, reaching up to 10 cm accuracy with 1000 Hz refresh rates. As a consequence, UWB is a popular technology for applications such as assets tracking in industrial environments or robots/drones indoor navigation. However, some security flaws in 802.15.4 standard expose UWB positioning to attacks. First, we showcase an exploit that remotely takes control of the pump. Then, we demonstrate the effectiveness of BackFlow, an LLVM-based compiler extension that enforces control-flow integrity in low-end ARM embedded systems.

More information ...

**DESIGN AUTOMATION FOR EXTENDED BURST-MODE AUTOMATA IN WORKCRAFT**

**Authors:**
Alex Chan, Alex Yakovlev, Danil Sokolov and Victor Khomenko, Newcastle University, GB

Abstract

Asynchronous circuits are known to have high performance, robustness and low power consumption, which are particularly beneficial for the area of so-called “little digital” controllers where low latency is crucial. However, asynchronous design is not widely adopted by industry, partially due to the steep learning curve inherent in the complexity of formal specifications, such as Signal Transition Graphs (STGs). In this demo, we promote a class of the Finite State Machine (FSM) model called Extended Burst-Mode (XBM) automata as a practical way to specify many asynchronous circuits. The XBM specification has been automated in the Workcraft toolkit (https://workcraft.org) with elaborate support for state encoding, conditional and “don’t care” signals. Formal verification and logic synthesis of the XBM automata is implemented via conversion to the established STG model, reusing existing methods and CAD tools. Tool support for the XBM flow will be demonstrated using several case studies.

More information ...
This session focuses on the advantages and use of novel emerging nanotechnology devices and their use in designing circuits and systems for embedded AI hardware solutions.

**More information ...**

**UB05.8** CATANIS: CAD TOOL FOR AUTOMATIC NETWORK SYNTHESIS

**Authors:**
Davide Quagliola, Enrico Fraccaroli, Filippo Nevi and Sohail Mushig, Università di Verona, IT

**Abstract**
The proliferation of communication technologies for embedded systems opened the way for new applications, e.g., Smart Cities and Industry 4.0. In such applications hundreds or thousands of smart devices interact together through different types of channels and protocols. This increasing communication complexity forces computer-aided design methodologies to scale up from embedded systems in isolation to the global inter-connected system. Network Synthesis is the methodology to optimally allocate functionality onto network nodes and define the communication infrastructure among them. This booth will demonstrate the functionality of a graphic tool for automatic network synthesis developed by the Computer Science Department of University of Verona. It allows to graphically specify the communication requirements of a smart space (e.g., its map can be considered) in terms of sensing and computation tasks together with a library of node types and communication protocols to be used.

**More information ...**

**UB05.9** PARALLEL ALGORITHM FOR CNN INFERENCE AND ITS AUTOMATIC SYNTHESIS

**Authors:**
Takashi Matsumoto, Yukio Miyasaka, Xinpei Zhang and Masahiro Fujita, University of Tokyo, JP

**Abstract**
Recently, Convolutional Neural Network (CNN) has surpassed conventional methods in the field of image processing. This demonstration shows a new algorithm to calculate CNN inference using processing elements arranged and connected based on the topology of the convolution. They are connected in mesh and calculate CNN inference in a systolic way. The algorithm performs the convolution of all elements with the same output feature in parallel. We demonstrate a method to automatically synthesize an algorithm, which simultaneously performs the convolution and the communication of pixels for the computation of the next layer. We show with several sizes of input layers, kernels, and strides and confirmed that the correct algorithms were synthesized. The synthesis method is extended to the sparse kernel. The synthesized algorithm requires fewer cycles than the original algorithm. There were the more chances to reduce the number of cycles with the sparser kernel.

**More information ...**

**UB05.10** FU: LOW POWER AND ACCURACY CONFIGURABLE APPROXIMATE ARITHMETIC UNITS

**Authors:**
Tomooi Ukezono and Toshinori Sato, Fukuoka University, JP

**Abstract**
In this demonstration, we will introduce the approximate arithmetic units such as adder, multiplier, and MAC that are being studied in our system-architecture laboratory. Our approximate arithmetic units can reduce delay and power consumption at the expense of accuracy. Our approximate arithmetic units are intended to be applied to IoT edge devices that can process images, and are suitable for battery-driven and low-cost devices. The biggest feature of our approximate arithmetic units is that the circuit is configured so that the accuracy is dynamically variable, and the trade-off relationship between accuracy and power can be selected according to the usage status of the device. In this demonstration, we show the power consumption according to various relationship between accuracy and power can be selected automatically synthesize an algorithm, which simultaneously performs the convolution and the communication of pixels for the computation of the next layer. We show with several sizes of input layers, kernels, and strides and confirmed that the correct algorithms were synthesized. The synthesis method is extended to the sparse kernel. The synthesized algorithm requires fewer cycles than the original algorithm. There were the more chances to reduce the number of cycles with the sparser kernel.

**More information ...**

12:00 End of session

**6.1 Special Day on “Embedded AI”: Emerging Devices, Circuits and Systems**

**Date:** Wednesday 11 March 2020

**Time:** 11:00 - 12:30

**Location / Room:** Amphithéâtre Jean Prouve

**Chair:**
Carlo Reita, CEA, FR

**Co-Chair:**
Bernabe Linares-Barranco, CSIC, ES

This session focuses on the advantages and use of novel emerging nanotechnology devices and their use in designing circuits and systems for embedded AI hardware solutions.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00</td>
<td>6.1.1</td>
<td>IN-MEMORY RESISTIVE RAM IMPLEMENTATION OF BINARIZED NEURAL NETWORKS FOR MEDICAL APPLICATIONS</td>
<td>Damien Querlioz, University Paris-Saclay, FR; Bogdan Penkovsky,1 Marc Bouquet,2 Tifenn Hirtzlin,1 Jacques-Olivier Klein,1 Etienne Nowak,1 Elisa Vianello,3 Jean-Michel Portal2 and Damien Querlioz4</td>
</tr>
</tbody>
</table>

1Université Paris-Saclay, FR; 2Aix-Marseille University, FR; 3CEA-Leti, FR; 4Université Paris-Sud, FR

**Abstract**
The advent of deep learning has considerably accelerated machine learning development, but its development at the edge is limited by its high energy cost and memory requirement. With new memory technology available, emerging Binarized Neural Networks (BNNs) are promising to reduce the energy impact of the forthcoming machine learning hardware generation, enabling machine learning on the edge devices and avoiding data transfer over the network. In this work, after presenting our implementation employing a hybrid CMOS -hafnium oxide resistive memory technology, we suggest strategies to apply BNNs to biomedical signals such as electrocardiography and electroencephalography, keeping accuracy level and reducing memory requirements. These results are obtained in binarizing solely the classifier part of a neural network. We also discuss how these results translate to the edge-oriented Mobilenet V1 neural network on the Imagenet task. The final goal of this research is to enable smart autonomous healthcare devices.

Download Paper (PDF; Only available from the DATE venue WiFi)
specific techniques range from the designs of integrity trees and hash tables, the management of superpages in filesystems, data prefetch in solid state drives (SSDs), as well as

As memories become persistent, the design of traditional data structures such as trees and hash tables as well as filesystems should be revisited to cope with the challenges brought by new memory devices. In this context, the main focus of this session is on how to improve performance, security, and energy-efficiency of memory and storage. The large capacity, the SGX-style integrity tree (SIT) is practical due to its parallel updates and variable arity. However, employing SIT in secure NVM is not easy. This is because the secure metadata SIT must be strictly persisted or restored after a sudden power-loss, which unfortunately incurs unacceptable run-time overhead or recovery time. In this paper, we propose PSIT, a metadata persistency solution for SIT-protected secure NVM with high performance and fast restoration. PSIT utilizes the observation that for a lazily updated SIT, the lost tree nodes after a crash can be recovered by the corresponding child nodes in the NVM. It reduces the persistency overhead of the SIT nodes through a restrained write-back meta-cache and leverages the SIT inter-layer dependency for recovery. Experiments show that compared to ASIT, a state-of-the-art secure NVM using SIT, PSIT decreases write traffic by 47% and improves the performance by 18% on average while maintaining a comparable recovery time.

Download Paper (PDF; Only available from the DATE venue WiFi)
Emerging non-volatile memory technologies bring evolution to storage systems and durable data structures. Among them, a proliferation of researches on persistent hash table employ NVM as the storage layer for both fast access and efficient persistence. Most of them are based on the assumptions that NVM has byte access granularity, poor write endurance, DRAM-comparable read latency and much higher write latency. However, a commercial non-volatile memory product, named Intel Optane DC Persistent Memory (AEP), has a few interesting features that are different from previous assumptions, such as 1) block access granularity, 2) little concern for software-layer write endurance and 3) much higher read latency than DRAM and DRAM-comparable write latency. Confronted with the new challenges brought by AEP, we propose Rewo-Hash, a novel read-efficient and write-optimized hash table for commercial non-volatile memory. Our design can be summarized into three key points. First, we keep a hash table copy in DRAM as a cached table to speed up search requests. Second, we design a log-free atomic mechanism to support fast writes. Third, we devise an efficient synchronization scheme between the persistent table and cached table to mask the data synchronization overhead. We conduct extensive experiments using real NVM platform and the results show that compared with state-of-the-art NVM-Optimized hash tables, Rewo-Hash gains improvement of 1.73x-2.70x and 1.46x-3.11x in read latency and write latency, respectively. Rewo-Hash also outperforms its counterparts by 1.86x-4.24x in throughput for various YCSB workloads.

Download Paper (PDF; Only available from the DATE venue WiFi)
potential location for rectification and any node in the logic network may be used to simplify the synthesized patch. Furthermore, off-the-self QBF algorithms do not allow a formulation of resource costs for reusing existing logic.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00</td>
<td>6.3.1</td>
<td>ENGINEERING CHANGE ORDER FOR COMBINATIONAL AND SEQUENTIAL DESIGN RECTIFICATION</td>
<td>Speaker: Jie-Hong Roland Jiang, National Taiwan University, TW Authors: Jie-Hong Roland Jiang¹, Victor Kravets² and Nian-Ze Lee¹ ¹National Taiwan University, TW; ²IBM, US</td>
</tr>
<tr>
<td>11:20</td>
<td>6.3.2</td>
<td>EXACT DAG-AWARE REWRITING</td>
<td>Speaker: Heinz Rieger, EPFL, CH Authors: Heinz Rieger¹, Alan Mishchenko² and Mathias Soeken¹ ¹EPFL, CH; ²University of California, Berkeley, US</td>
</tr>
<tr>
<td>11:40</td>
<td>6.3.3</td>
<td>LEARNING TO AUTOMATE THE DESIGN UPDATES FROM OBSERVED ENGINEERING CHANGES IN THE CHIP DEVELOPMENT CYCLE</td>
<td>Speaker: Victor Kravets, IBM, US Authors: Victor Kravets¹, Jie-Hong Roland Jiang² and Heinz Rieger³ ¹IBM, US; ²National Taiwan University, TW; ³EPFL, CH</td>
</tr>
<tr>
<td>12:05</td>
<td>6.3.4</td>
<td>SYNTHESIS AND OPTIMIZATION OF MULTIPLE PORTIONS OF CIRCUITS FOR ECO BASED ON SET-COVERING AND QBF FORMULATIONS</td>
<td>Speaker: Masahiro Fujita, University of Tokyo, JP Authors: Masahiro Fujita, Yusuke Kimura, Xingming Le, Yukio Miyasaka and Amir Masoud Ghareshbaghi, University of Tokyo, JP</td>
</tr>
</tbody>
</table>

**6.4 Microarchitecture to the rescue of memory**

**Date:** Wednesday 11 March 2020  
**Time:** 11:00 - 12:30  
**Location / Room:** Stendhal  
**Chair:** Olivier Sentieys, INRIA, FR  
**Co-Chair:** Jeronimo Castrillon, TU Dresden, DE

This session discusses micro-architectural innovations across three different memory technologies, namely, caches, 3D-stacked DRAM and non-volatile. This includes exploiting several aspects of redundancy to maximize cache utilization through compression, as well as multicast in 3D-stacked high-speed memories for graph analytics, and a microarchitecture solution to unify persistency and encryption in non-volatile memories.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00</td>
<td>6.4.1</td>
<td>EFFICIENT HARDWARE-ASSISTED CRASH CONSISTENCY IN ENCRYPTED PERSISTENT MEMORY</td>
<td>Speaker: Zhan Zhang, Huazhong University of Science &amp; Technology, CN Authors: Zhan Zhang¹, Jianhui Yue², Xiaofei Liao³ and Hai Jin¹ ¹Huazhong University of Science &amp; Technology, CN; ²Michigan Technological University, US</td>
</tr>
</tbody>
</table>

**Abstract**

The persistent memory (PM) requires maintaining the crash consistency and encrypting data, to ensure data recoverability and data confidentiality. The enforcement of these two goals does not only put more burden on programmers but also degrades performance. To address this issue, we propose a hardware-assisted encrypted persistent memory system. Specifically, logging and data encryption are assisted by hardware. Furthermore, we apply the counter-based encryption and the cipher feedback (CFB) mode encryption to data and log respectively, reducing the encryption overhead. Our primary experimental results show that the transaction throughput of the proposed design outperforms the baseline design by up to 34.4%.

Download Paper (PDF; Only available from the DATE venue WiFi)
Sebastian Steinhorst, TU Munich, DE

Brandon Reagen, Facebook and New York University, US

Chair:

Saveen Reddy, Facebook, US

Location / Room:

Bayard

The large processing requirements of ML models strain the capabilities of low-power embedded systems. Addressing this challenge, the first presentation proposes a robust co-design to leverage stochastic computing for highly accurate and efficient inference. Next, a structural optimization is proposed to counter faults at low voltage levels. Then, authors present a method for sharing results in binarized CNNs to reduce computation. The session will conclude with a talk implementing binary networks on mobile GPUs.
ACOUSTIC: ACCELERATING CONVOLUTIONAL NEURAL NETWORKS THROUGH OR-UNIPOLAR SKIPPED STOCHASTIC COMPUTING

Speaker: Puneet Gupta, University of California, Los Angeles, US
Authors: Wojciech Romaszkan, Tianmu Li, Tristan Melton, Sudhakar Pamarti and Puneet Gupta, University of California, Los Angeles, US

Abstract
As privacy and latency requirements force a move towards edge Machine Learning inference, resource constrained devices are struggling to cope with large and computationally complex models. For Convolutional Neural Networks, those limitations can be overcome by taking advantage of enormous data reuse opportunities and amenability to reduced precision. To do that however, a level of compute density unattainable for conventional binary arithmetic is required. Stochastic Computing can deliver such density, but it has not lived up to its full potential because of multiple underlying precision issues. We present ACOUSTIC: Accelerating Convolutions through Or-Unipolar Skipped Stochastic Computing, an accelerator framework that enables fully stochastic, high-density CNN inference. Leveraging split-unipolar representation, OR-based accumulation and novel computation-skipping approach, ACOUSTIC delivers server-class parallelism within a mobile area and power budget - a 12mm2 accelerator can be as much as 38.7x more energy efficient and 72.5x faster than conventional fixed-point accelerators. It can also be up to 79.6x more energy efficient than state-of-the-art stochastic accelerators. At the lower-end ACOUSTIC achieves 8x-120x inference throughput improvement with similar energy and area when compared to recent mixed-signal/neuromorphic accelerators.

Download Paper (PDF; Only available from the DATE venue WiFi)

ACCURACY TOLERANT NEURAL NETWORKS UNDER AGGRESSIVE POWER OPTIMIZATION

Speaker: Yi-Wen Hung, National Tsing Hua University, TW
Authors: Xiang-Xiu Wu1, Yi-Wen Hung1, Yung-Chih Chen2 and Shih-Chieh Chang1
1National Tsing Hua University, TW; 2Yuan Ze University, Taoyuan, Taiwan, TW

Abstract
With the success of deep learning, many neural network models have been proposed and applied to various applications. In several applications, the devices used to implement the complicated models have limited power resources and thus aggressive optimization techniques are often applied for saving power. However, some optimization techniques, such as voltage scaling and multiple threshold voltages, may increase the probability of error occurrence due to slow signal propagation, which increases the path delay in a circuit and fails some input patterns. Although neural network models are considered to have some error tolerance, the prediction accuracy could be significantly affected, when there are a large number of errors. Thus, in this paper, we propose a scheme to mitigate the errors caused by slow signal propagation. Since the delay of multipliers dominate the critical path of the circuit, we consider the patterns significantly altered by the slow signal propagation according to the multipliers and prevent the patterns from failure by adjusting the neural network and the parameters. The proposed scheme modifies a neural network on the software side and thus it is unnecessary to re-design the hardware structure. The experimental results show that the proposed scheme is effective for several neural network models. It can improve the accuracy up to 27%, when the device under consideration is applied with aggressive power optimization techniques.

Download Paper (PDF; Only available from the DATE venue WiFi)

A CONVOLUTIONAL RESULT SHARING APPROACH FOR BINARIZED NEURAL NETWORK INFERENCE

Speaker: Chia-Chun Lin, National Tsing Hua University, TW
Authors: Ya-Chun Chang1, Chia-Chun Lin1, Yi-Ting Lin1, Yung-Chih Chen1 and Chun-Yao Wang1
1National Tsing Hua University, TW; 2Yuan Ze University, Taoyuan, Taiwan, TW

Abstract
The binary-weight-binary-input binarized neural network (BNN) allows a much more efficient way to implement convolutional neural networks (CNNs) on mobile platforms. During inference, the multiply-accumulate operations in BNNs can be reduced to XNOR-popcount operations. Thus, the XNOR-popcount operations dominate most of the computation in BNNs. To reduce the number of required operations in convolution layers of BNNs, we decompose 3-D filters into 2-D filters and exploit the repeated filters, inverse filters, and similar filters to share results. By sharing the results, the number of operations in convolution layers of BNNs can be reduced effectively. Experimental results show that the number of operations can be reduced by about 60% for CIFAR-10 on BNNs while keeping the accuracy loss within 1% of originally trained network.

Download Paper (PDF; Only available from the DATE venue WiFi)

PHONEDBIT: EFFICIENT GPU-ACCELERATED BINARY NEURAL NETWORK INFERENCE ENGINE FOR MOBILE PHONES

Speaker: Gang Chen, Sun Yat-sen University, CN
Authors: Gang Chen1, Shengyu He1, Haitao Meng2 and Kai Huang1
1Sun Yat-sen University, CHN; 2Northeastern University, CN

Abstract
Over the last years, a great success of deep neural networks (DNNs) has been witnessed in computer vision and other fields. However, performance and power constraints make it still challenging to deploy DNNs on mobile devices due to their high computational complexity. Binary neural networks (BNNs) have been demonstrated as a promising solution to achieve this goal by using bit-wise operations to replace most arithmetic operations. Currently, existing GPU-accelerated implementations of BNNs are only tailored for desktop platforms. Due to architecture differences, mere porting of such implementations to mobile devices yields suboptimal performance or is impossible in some cases. In this paper, we propose PhonedBit, a GPU-accelerated BNN inference engine for Android-based mobile devices that fully exploits the computing power of BNNs on mobile GPUs. PhonedBit provides a set of operator-level optimizations including locality-friendly data layout, bit packing with vectorization and layers integration for efficient binary convolution. We also provide a detailed implementation and parallelization optimization for PhonedBit to optimally utilize the memory bandwidth and computing power of mobile GPUs. We evaluate PhonedBit with AlexNet, YOLOv2 Tiny and VGG16 with their binary version. Our experimental results show that PhonedBit can achieve significant speedup and energy efficiency compared with state-of-the-art frameworks for mobile devices.

Download Paper (PDF; Only available from the DATE venue WiFi)

HARDWARE ACCELERATION OF CNN WITH ONE-HOT QUANTIZATION OF WEIGHTS AND ACTIVATIONS

Speaker: Gang Li, Chinese Academy of Sciences, CN
Authors: Gang Li, Peisong Wang, Zejian Liu, Cong Leng and Jian Cheng, Chinese Academy of Sciences, CN

Abstract
In this paper, we propose a novel one-hot representation for weights and activations in CNN model and demonstrate its benefits on hardware accelerator design. Specifically, rather than merely reducing the bitwidth, we quantize both weights and activations into n-bit integers that containing only one non-zero bit per value. In this way, the massive multiply and accumulates (MACs) are equivalent to additions of powers of two that can be efficiently calculated with histogram based computations. Experiments on the ImageNet classification task show that comparable accuracy can be obtained on our proposed One-Hot Networks (OHN) compared to conventional fixed-point networks. As case studies, we evaluate the efficacy of one-hot data representation on two state-of-the-art CNN accelerators on FPGA, our preliminary results show that 50% and 68.5% resource saving can be achieved on DaDianNao and Laconic respectively. Besides, the one-hot optimized Laconic can further achieve an average speedup of 4.94x on AlexNet.

Download Paper (PDF; Only available from the DATE venue WiFi)
6.6 From DFT to Yield Optimization

Date: Wednesday 11 March 2020
Time: 11:00 - 12:30
Location / Room: Lesdiguières
Chair: Maria Micheal, University of Cyprus, CY
Co-Chair: Sanchez Ernesto, Politecnico di Torino, IT

The session presents a variety of semiconductor test techniques, including a new design-for-testability scheme for FinFET SRAMs, a method to increase yield based on error-tolerant RSNs, and a new algorithm to compute scan paths in faulty RSNs. The session also includes a discussion on fault-tolerant applications in distributed systems.

12:31 IP3-4, 729 A DFT SCHEME TO IMPROVE COVERAGE OF HARD-TO-DETECT FAULTS IN FINFET SRAMS
Speaker: Guilherme Cardoso Medeiros, TU Delft, NL
Authors: Guilherme Cardoso Medeiros 1, Cemil Cem Gürsoy 2, Moritz Fleibach 1, Lihou Wu 1, Maksim Jeninhiin 2, Mottaqiallah Taouil 1 and Said Hamdioui 1
1TU Delft, NL; 2Tallinn University of Technology, EE
Abstract: Manufacturing defects can cause faults in FinFET SRAMs. Of them, easy-to-detect (ETD) faults always cause incorrect behavior, and therefore are easily detected by applying sequences of write and read operations. However, hard-to-detect (HTD) faults may not cause incorrect behavior, only parametric deviations. Detection of these faults is of major importance as they may lead to test escapes. This paper proposes a new design-for-testability (DFT) scheme for FinFET SRAMs to detect such faults by creating a mismatch in the sense amplifier (SA). This mismatch, combined with the defect in the cell, will incorrectly bias the SA and cause incorrect read outputs. Furthermore, post-silicon calibration schemes can be used to avoid over-testing or test escapes caused by process variation effects. Compared to the state of the art, this scheme introduces negligible overheads in area and test time while it significantly improves fault coverage and reduces the number of test escapes.
Download Paper (PDF; Only available from the DATE venue WiFi)

12:32 IP3-5, 147 A HIGHLY ACCURATE LOG_2 LEAD QUANTIZATION OF PRE-TRAINED NEURAL NETWORKS
Speaker: Salim Ulah, TU Dresden, DE
Authors: Salim Ulah1, Siddharta Gupta2, Kapil Ahuja2, Aruna Tiwari2 and Akash Kumar1
1TU Dresden, DE; 2IIT Indore, IN
Abstract: Deep Neural Networks are one of the machine learning techniques which are increasingly used in a variety of applications. However, the significantly high memory and computation demands of deep neural networks often limit their deployment on embedded systems. Many recent works have considered this problem by proposing different types of data quantization schemes. However, most of these techniques either require post-quantization retraining of deep neural networks or bear a significant loss in output accuracy. In this paper, we propose a novel quantization technique for parameters of pre-trained deep neural networks. Our technique significantly maintains the accuracy of the parameters and does not require retraining of the networks. Compared to the single-precision floating-point numbers-based implementation, our proposed 8-bit quantization technique generates only –1% and –0.6% loss in the top-1 and top-5 accuracies respectively for VGG16 network using ImageNet dataset.
Download Paper (PDF; Only available from the DATE venue WiFi)

End of session
6.7 Safety and efficiency for smart automotive and energy systems

Date: Wednesday 11 March 2020
Time: 11:00 - 12:30
Location / Room: Berioz

Chair: Selma Saidi, TU Dortmund, DE
Co-Chair: Donghwa Shin, Soongsil University, KR

This session presents four papers dealing with various aspects of smart automotive and energy systems, including safety and efficiency of photovoltaic panels, deterministic behavior execution of adaptive automotive applications, efficient implementation of fail-operational automated vehicles, and efficient resource usage in networked automotive systems.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00</td>
<td>6.6.3</td>
<td>USING PROGRAMMABLE DELAY MONITORS FOR WEAR-OUT AND EARLY LIFE FAILURE PREDICTION</td>
<td>Speaker: Chang Liu, Altan Deutschland, DE Authors: Chang Liu, Eric Schneider and Hans-Joachim Wunderlich, University of Stuttgart, DE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abstract</td>
<td>Early life failures in marginal devices are a severe reliability threat in current nano-scaled CMOS devices. While small delay faults are an effective indicator of marginalities, their detection requires special efforts in testing by so-called Faster-than-At-Speed Test (FAST). In a similar way, delay degradation is an indicator that a device reaches the wear-out phase due to aging. Programmable delay monitors provide the possibility to detect gradual performance changes in a system and allow to observe device degradation. This paper presents a unified approach to test small delay faults related to wear-out and early-life failures by reuse of existing programmable delay monitors within FAST. The approach is complemented by a test-scheduling which optimally selects frequencies and delay configurations to significantly increase the fault coverage of small delays and to reduce the test time. Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
<tr>
<td>12:15</td>
<td>6.6.4</td>
<td>MAXIMIZING YIELD FOR APPROXIMATE INTEGRATED CIRCUITS</td>
<td>Speaker: Marcello Triolita, Université de Montpellier, FR Authors: Marcello Triolita¹, Arnaud Virazel², Patrick Girard², Mario Barbareschi³ and Alberto Bosio⁴ ¹LIRMM, FR; ²LIRMM / CNRS, FR; ³Università di Napoli Federico II, IT; ⁴Lyon Institute of Nanotechnology, FR Abstract Approximate Integrated Circuits (AxICs) have emerged in the last decade as an outcome of Approximate Computing (AxC) paradigm. AxC focuses on efficiency of computing systems by sacrificing some computation quality. As AxICs spread, consequent challenges to test them arose. On the other hand, the opportunity to increase the production yield emerged in the AxIC context. Indeed, some particular defects in the manufactured AxIC might not catastrophically impact the final circuit quality. Therefore, some defective AxICs might still be acceptable. Efforts to detect favorable conditions to consider defective AxICs as acceptable - with the goal to increase the production yield - have been done in last years. Unfortunately, the final achieved yield gain is often not as high as expected. In this work, we propose a methodology to actually achieve a yield gain as close as possible to expectations, by proposing a technique to suitably apply tests to AxICs. Experiments carried out on state-of-the-art AxICs show yield gain results very close to the expected ones (i.e., between 98% and 100% of the expectations). Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
<tr>
<td>12:30</td>
<td>IP3-6,359</td>
<td>FAULT DIAGNOSIS OF VIA-SWITCH CROSSBAR IN NON-VOLATILE FPGA</td>
<td>Speaker: Ryutaro Doi, Osaka University, JP Authors: Ryutaro Doi¹, Xu Bai², Toshtitsu Sakamoto³ and Masanori Hashimoto¹ ¹Osaka University, JP; ²NEC Corporation, Japan Abstract FPGA that exploits via-switches, which are a kind of non-volatile resistive RAMs, for crossbar implementation is attracting attention due to its high integration density and energy efficiency. Via-switch crossbar is responsible for the signal routing by changing on/off-states of via-switches. To verify the via-switch crossbar functionality after manufacturing, fault testing that checks whether we can turn on/off via-switches normally is essential. This paper confirms that a general differential pair comparator successfully discriminates on/off-states of via-switches, and clarifies fault modes of a via-switch by transistor-level SPICE simulation that injects stuck-on/off faults to atom switch and varistor, where a via-switch consists of two atom switches and two varistors. We then propose a fault diagnosis methodology that diagnoses the fault modes of each via-switch using the comparator response difference between normal and faulty via-switches. The proposed method achieves 100% fault detection by checking the comparator responses after turning on/off the via-switch. In case that the number of faulty components in a via-switch is one, the ratio of the fault diagnosis, which exactly identifies the faulty varistor and atom switch inside the faulty via-switch, is 100%, and in case of up to two faults, the fault diagnosis ratio is 79%. Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
<tr>
<td>12:30</td>
<td></td>
<td>End of session</td>
<td></td>
</tr>
</tbody>
</table>

6.7.1 A DIODE-AWARE MODEL OF PV MODULES FROM DATASHEET SPECIFICATIONS

Speaker: Sara Vinco, Politecnico di Torino, IT Authors: Sara Vinco, Yukai Chen, Enrico Macil and Massimo Poncino, Politecnico di Torino, IT Abstract Semi-empirical models of photovoltaic (PV) modules based on datasheet information are popular in electricalenergy systems (EES) simulation because they can be built without measurements and allow quick exploration of alternative devices. One key limitation of these models, however, is the fact that they cannot model the presence of bypass diodes, which are inserted across a set of series-connected cells in a PV module to mitigate the impact of partial shading; datasheet information refer in fact to the operations of the module under uniform irradiance. Neglecting the effect of bypass diodes may incur insignificant underestimation of the extracted power. This paper proposes a semi-empirical model for a PV module, that, by taking into account the only available information about bypass diodes in a datasheet, i.e., its number, by a first downsampling the model to a single PV cell and a subsequent upsampling to the level of a substring and of a module, allows to take into account the diode effect as much accurately as allowed by the datasheet information. Experimental results show that, in a typical PV array on a roof, using a diode-agnostic model can significantly underestimate the output power production. Download Paper (PDF; Only available from the DATE venue WiFi)
### ACHIEVING DETERMINISM IN ADAPTIVE AUTOSAR

**Speaker:** Christian Menard, TU Dresden, DE  
**Authors:** Christian Menard¹, Andres Goens¹, Marten Lohstroh² and Jeronimo Castrillon¹  
¹TU Dresden, DE; ²University of California, Berkeley, US  

**Abstract**

AUTOSAR Adaptive Platform is an emerging industry standard that tackles the challenges of modern automotive software design, but does not provide adequate mechanisms to enforce deterministic execution. This poses profound challenges to testing and maintenance of the application software, which is particularly problematic for safety-critical applications. In this paper, we analyze the problem of nondeterminism in AP and propose a framework for the design of deterministic automotive software that transparently integrates with the AP communication mechanisms. We illustrate our approach in a case study based on the brake assistant demonstrator application that is provided by the AUTOSAR consortium. We show that the original implementation is nondeterministic and discuss a deterministic solution based on our framework.

Download Paper (PDF; Only available from the DATE venue WiFi)

### A FAIL-SAFE ARCHITECTURE FOR AUTOMATED DRIVING

**Speaker:** Sebastian vom Dorf, DENSO Automotive Deutschland GmbH, DE  
**Authors:** Sebastian vom Dorf¹, Bert Böddeker², Maximilian Kneissl¹ and Martin Fränzle³  
¹DENSO Automotive Deutschland GmbH, DE; ²Autonomous Intelligent Driving GmbH, DE; ³Carnegie Mellon University Oldenburg, DE  

**Abstract**

The development of autonomous vehicles has gained a rapid pace. Along with the promising possibilities of such automated systems, the question of how to ensure their safety arises. With increasing levels of automation the need for fail-operational systems, not relying on a back-up driver, poses new challenges in system design. In this paper we propose a lightweight architecture addressing the challenge of a verifiable, fail-safe safety implementation for trajectory planning. It offers a distributed design and the ability to comply with the requirements of ISO26262, without an overely redundant set-up. Furthermore, we show an example with low-level prediction models applied to a real world situation.

Download Paper (PDF; Only available from the DATE venue WiFi)

### PRIORITY-PRESERVING OPTIMIZATION OF STATUS QUO ID-ASSIGNMENTS IN CONTROLLER AREA NETWORK

**Speaker:** Lea Schönberger, TU Dortmund University, DE  
**Authors:** Sebastian Schwitalia¹, Lea Schönberger¹ and Jian-Jia Chen²  
¹TU Dortmund University, DE; ²TU Dortmund, DE  

**Abstract**

Controller Area Network (CAN) is the prevailing solution for connecting multiple electronic control units (ECUs) in automotive systems. Every broadcast message on the bus is received by each bus participant and introduces computational overhead to the typically resource-constrained ECUs due to interrupt handling. To reduce this overhead, hardware message filters can be applied. However, since such filters are configured according to the message identifiers (IDs) specified in the system, the filter quality is limited by the nature of the ID-assignment. Although hardware message filters are highly relevant for industrial applications, so far, only the optimization of the filter design, but not the related optimization of ID-assignments has been addressed in the literature. In this work, we explicitly focus on the optimization of message ID-assignments against the background of hardware message filtering. More precisely, we propose an optimization algorithm transforming a given ID-assignment in such a way that, based on the resulting IDs, the quality of hardware message filters is improved significantly, i.e., the computational overhead introduced to each ECU is minimized, and, moreover, the priority order of the system remains unchanged. Conducting comprehensive experiments on automotive benchmarks, we show that our proposed algorithm clearly outperforms optimizations based on the conventional method simulated annealing with respect to the achieved filter quality as well as to the runtime.

Download Paper (PDF; Only available from the DATE venue WiFi)

### APPLYING RESERVATION-BASED SCHEDULING TO A µC-BASED HYPERVISOR: AN INDUSTRIAL CASE STUDY

**Speaker:** Dirk Ziegenbein, Robert Bosch GmbH, DE  
**Authors:** Dakshina Dasari¹, Paul Austin², Michael Pressler¹, Arne Hamann¹ and Dirk Ziegenbein¹  
¹Robert Bosch GmbH, DE; ²ETAS GmbH, GB  

**Abstract**

Existing software scheduling mechanisms do not suffice for emerging applications in the automotive space, which have the conflicting needs of performance and predictability. We need mechanisms that lend themselves naturally to this requirement, by virtue of their design. As a concrete case, we consider the ETAS light-weight hypervisor, a commercially viable solution in the automotive industry, deployed on multicore microcontrollers. We describe the architecture of the hypervisor and its current scheduling mechanisms based on Time Division Multiplexing. We next show how Reservation-based Scheduling can be implemented in the ETAS LWVHR to efficiently use resources while also providing freedom from interference and explore design choices towards an efficient implementation of such a scheduler. With experiments from an industry use case, we also compare the performance of RBS and the existing scheduler in the hypervisor.

Download Paper (PDF; Only available from the DATE venue WiFi)

### REAL-TIME ENERGY MONITORING IN IOT-ENABLED MOBILE DEVICES

**Speaker:** Nitin Shivaraman, TUMCREATE, SG  
**Authors:** Nitin Shivaraman¹, Seima Surjyasekaran¹, Zhiwei Liu², Saravanan Ramanathan¹, Arvind Easwaran² and Sebastian Steinhorst³  
¹TUMCREATE, SG; ²Nanyang Technological University, SG; ³TU Munich, DE  

**Abstract**

With rapid advancements in the Internet of Things (IoT) paradigm, every electrical device in the near future is expected to have IoT capabilities. This enables fine-grained tracking of individual energy consumption data of such devices, offering location-independent per-device billing and demand management. Hence, it abstracts from the location-based metering of state-of-the-art infrastructure, which traditionally aggregates on a building or household level, defining the entity to be billed. However, such in-device energy monitoring is susceptible to manipulation and fraud. As a remedy, we propose a secure decentralized metering architecture that enables devices with IoT capabilities to measure their own energy consumption. In this architecture, the device-level consumption is additionally reported to a system-level aggregator that verifies distributed information from our decentralized metering systems and provides secure data storage using Blockchain, preventing data manipulation by untrusted entities. Through experimental evaluation, we show that the proposed architecture supports device mobility and enables location-independent monitoring of energy consumption.

Download Paper (PDF; Only available from the DATE venue WiFi)
A DIGITAL MICROFLUIDICS BIO-COMPUTING PLATFORM
Authors: Tharam Selvam, Luca Pezzarossa, Winnie Edith Svendsen and Jan Madsen, TU Denmark, DK
Abstract: Digital microfluidics is a lab-on-a-chip (LOC) technology used to actuate small amounts of liquids on an array of individually addressable electrodes. Microliter sized droplets can be programmatically dispensed, moved, mixed, split, in a controlled environment which combined with miniaturized sensing techniques makes LOC suitable for a broad range of applications in the field of medical diagnostics and synthetic biology. Furthermore, a programmable digital microfluidics platform holds the potential to add to a "fluidic subsystem" to the classical computation model thus opening the doors for cyber-physical bio-processors. To facilitate the programming and operation of such bio-fluidic computing, we propose dedicated instruction set architecture and virtual machine. A set of digital microfluidic core instructions as well as classic computing operations are executed on a virtual machine, which decouples the protocol execution from the LOC functionality.

VIRTUAL PLATFORMS FOR COMPLEX SOFTWARE STACKS
Authors: Lukas Jünger and Rainer Leupers, RWTH Aachen University, DE
Abstract: This demonstration is going to showcase our "AVPG4" Virtual Platform (VP), which models a multi-core ARMv8 (Cortex A72) system including several peripherals, such as an SDHCI and an ethernet controller. For the ARMv8 instruction set simulation a dynamic binary translation based solution is used. As the workload, the Xen hypervisor with two Linux Virtual Machines (VMs) is executed. Both VMs are connected to the simulation host's network subsystem via a virtual ethernet controller. One of the VMs executes a NodeJS-based server application offering a REST API via this network connection. An AngularJS client application on the host system can then connect to the server application to obtain and store data via the server's REST API. This data is read and written by the server application to the virtual SD Card connected to the SDHCI. For this, one SD card partition is passed to the VM through Xen's block device virtualization mechanism.

SYSTEMC-CT/DE: A SIMULATOR WITH FAST AND ACCURATE CONTINUOUS TIME AND DISCRETE EVENTS INTERACTIONS ON TOP OF SYSTEMC.
Authors: Breytnner Joseph Fernandez-Mesa, Liliana Andrade and Frédéric Pétrot, Université Grenoble Alpes / CNRS / TIMA Laboratory, FR
Abstract: We have developed a continuous time (CT) and discrete events (DE) simulator on top of SystemC. Systems that mix both domains are critical and their proper functioning must be verified. Simulation serves to achieve this goal. Our simulator implements direct CT/DE synchronization, which enables a rich set of interactions between the domains: events from the CT models are able to trigger DE processes; events from the DE models are able to modify the CT equations. DE-based interactions are, then, simulated at their precise time by the DE kernel rather than at fixed time steps. We demonstrate our simulator by executing a set of challenging examples: they either require a superset model of time or include Zeno behavior or are highly sensitive to accuracy errors. Results show that our simulator overcomes these issues, is accurate, and improves simulation speed w.r.t. fixed time steps; all of these advantages open up new possibilities for the design of a wider set of heterogeneous systems.

SRSN: SECURE RECONFIGURABLE TEST NETWORK
Authors: Vincent Reynaud1, Emanuele Valea1, Paolo Maitri1, Regis Leveugle1, Marie-Lise Flottes1, Sophie Dupuis2, Bruno Rouzeyre2 and Giorgio Di Natale1
1TIMA Laboratory, FR; 2LIRMM, FR
Abstract: The critical importance of testability for electronic devices led to the development of IEEE test standards. These methods, if not protected, offer a security backdoor to attackers. This demonstrator illustrates a state-of-the-art solution that prevents unauthorized usage of the test infrastructure based on the IEEE 1687 standard and implemented on an FPGA target.

GENERATING ASYNCHRONOUS CIRCUITS FROM CATAPULT
Authors: Yoan Decoudou1, Jean Simatic2, Katell Morin-Allory3 and Laurent Fesquet3
1University Grenoble Alpes, FR; 2HavaAL.Tech, FR; 3Université Grenoble Alpes, FR
Abstract: In order to spread asynchronous circuit design to a large community of designers, High-Level Synthesis (HLS) is probably a good choice because it requires limited design technical skills. HLS usually provides an RTL description, which includes a data-path and a control-path. The desynchronization process is only applied to the control-path, which is a Finite State Machine (FSM). This method is sufficient to make asynchronous the circuit. Indeed, data are processed step by step in the pipeline stages, thanks to the desynchronized FSM. Thus, the data-path computation time is no longer related to the clock period but rather to the average time for processing data into the pipeline. This tends to improve speed when the pipeline stages are not well-balanced. Moreover, our approach helps to quickly design data-driven circuits while maintaining a reasonable cost, a similar area and a short time-to-market.

LEARNL: LEARNL: A RISC-V BASED EMBEDDED SYSTEM DESIGN FRAMEWORK FOR EDUCATION AND RESEARCH DEVELOPMENT
Authors: Noureddine Ait Said and Mourin Benabdenbi, TIMA Laboratory, FR
Abstract: Designing a modern System on a Chip is based on the joint design of hardware and software (co-design). However, understanding the tight relationship between hardware and software is not straightforward. Moreover to validate new concepts in SoC design from the idea to the hardware implementation is time-consuming and often slowed by legacy issues (intellectual property of hardware blocks and expensive commercial tools). To overcome these issues we propose to use the open-source Rocket Chip environment for educational purposes, combined with the open-source LowRisc architecture to implement a custom SoC design on an FPGA board. The demonstration will present how students and engineers can take benefit from the environment to deepen their knowledge in HW and SW co-design. Using the LowRisc architecture, an image classification application based on the use of CNNs will serve as a demonstrator of the whole open-source hardware and software flow and will be mapped on a Nexys A7 FPGA board.
UB06.9 WALLANCE: AN ALTERNATIVE TO BLOCKCHAIN FOR IOT
Authors: Loïc Dalmasso, Florent Bruguet, Pascal Benoit and Achraf Lamlih, Université de Montpellier, FR
Abstract
Since the expansion of the Internet of Things (IoT), connected devices became smart and autonomous. Their exponentially increasing number and their use in many application domains result in a huge potential of cybersecurity threats. Taking into account the evolution of the IoT, security and interoperability are the main challenges, to ensure the reliability of the information. The blockchain technology provides a new approach to handle the trust in a decentralized network. However, current blockchain implementations cannot be used in IoT domain because of their huge need of computing power and storage utilization. This demonstrator presents a lightweight distributed ledger protocol dedicated to the IoT application, reducing the computing power and storage utilization, handling the scalability and ensuring the reliability of information.
More information ...

UB06.10 JOINTER: JOINING FLEXIBLE MONITORS WITH HETEROGENEOUS ARCHITECTURES
Authors: Giacomo Valente, Tiziana Fanni, Carlo Sau, Claudio Rubattu, Francesca Palumbo and Luigi Pomante
1Università degli Studi dell’Aquila, IT; 2Università degli Studi di Sassari, IT; 3Università degli Studi di Cagliari, IT
Abstract
As embedded systems grow more complex and shift toward heterogeneous architectures, understanding workload performance characteristics becomes increasingly difficult. In this regard, run-time monitoring systems can support an obtaining the desired visibility to characterize a system. This demo presents a framework that allows to develop complex heterogeneous architectures composed of programmable processors and dedicated accelerators on FPGA, together with customizable monitoring systems, keeping under control the introduced overhead. The whole development flow (and related prototypal EDA tools), that starts with the accelerators creation using a dataflow model, in parallel with the monitoring system customization using a library of elements, showing also the final joining, will be shown. Moreover, a comparison among different monitoring systems functionalities on different architectures developed on Zynq7000 SoC will be illustrated.
More information ...

14:00 End of session

7.0 LUNCHTIME KEYNOTE SESSION
Date: Wednesday 11 March 2020
Time: 13:45 - 14:20
Location / Room: Amphithéâtre Jean Prouvé
Chair: Bernabe Linares-Barranco, CSIC, ES
Co-Chair: Dmitri Strukov, University of California, Santa Barbara, US

Time Label Presentation Title Authors

13:45 7.0.0 CEDA LUNCHEON ANNOUNCEMENT
Author: David Atienza, EPFL, CH

13:50 7.0.1 LEVERAGING EMBEDDED INTELLIGENCE IN INDUSTRY: CHALLENGES AND OPPORTUNITIES
Author: Jim Tung, MathWorks Fellow, US
Abstract
The buzz about AI is deafening. Compelling applications are starting to emerge, dramatically changing the customer service that we experience, the marketing messages that we receive, and some systems we use. But, as organizations decide whether and how to incorporate AI in their systems and services, they must bring together new combinations of specialized knowledge, domain expertise, and business objectives. They must navigate through numerous choices - algorithms, processors, compute placement, data availability, architectural allocation, communications, and more. At the same time, they must keep their focus on the applications that will create compelling value for them. In this keynote, Jim Tung looks at the promising opportunities and practical challenges of building AI into our systems and services.

14:20 End of session

UB07 Session 7
Date: Wednesday 11 March 2020
Time: 14:00 - 16:00
Location / Room: Booth 11, Exhibition Area

Label Presentation Title Authors

UB07.1 DL PUF ENAU: DEEP LEARNING BASED PHYSICALLY UNCLONABLE FUNCTION ENROLLMENT AND AUTHENTIFICATION
Authors: Amir Alipour1, David Hely2, Vincent Beroulle2 and Giorgio Di Natale3
1Grenoble INP / LCIS, FR; 2Grenoble INP, FR; 3CNRS / Grenoble INP / TIMA, FR
Abstract
Physically Unclonable Functions (PUFs) have been addressed nowadays as a potential solution to improve the security in authentication and encryption process in Cyber Physical Systems. The research on PUF is actively growing due to its potential of being secure, easily implementable and expandable, using considerably less energy. To use PUF in common, the low level device Hardware Variation is captured per unit for device enrollment into a format called Challenge-Response Pair (CRP), and recaptured after device is deployed, and compared with the original for authentication. These enrollment + comparison functions can vary and be more data demanding for applications that demand robustness, and resilience to noise. In this demonstration, our aim is to show the potential of using Deep Learning for enrollment and authentication of PUF CRPs. Most importantly, during this demonstration, we will show how this method can save time and storage compared to other classical methods.
More information ...
UB07.2 BCFLEASAM: BACKFLOW: BACKWARD EDGE CONTROL FLOW ENFORCEMENT FOR LOW END ARM REAL-TIME SYSTEMS
Authors: Bresch Cyril 1, David Hély 1, Roman Lysecky 2 and Stephanie Chollet 1
1ICIS, FR; 2University of Arizona, US
Abstract
The C programming language is one of the most popular languages in embedded system programming. Indeed, C is efficient, lightweight and can easily meet high performance and deterministic real-time constraints. However, these assets come at a certain price. Indeed, C does not provide extra features for memory safety. As a result, attackers can easily exploit spatial memory vulnerabilities to hijack the execution flow of an application. The demonstration features a real-time connected infusion pump vulnerable to memory attacks. First, we showcase an exploit that remotely takes control of the pump. Then, we demonstrate the effectiveness of BackFlow, an LLVM-based compiler extension that enforces control-flow integrity in low-end ARM embedded systems.
More information ...

UB07.3 A BINARY TRANSLATION FRAMEWORK FOR AUTOMATED HARDWARE GENERATION
Authors: Nuno Paulino and João Canas Ferreira, INESC TEC / University of Porto, PT
Abstract
Hardware specialization is an efficient solution for maximization of performance and minimization of energy consumption. This work is based on automated detection of workload by analysis of a compiled application, and on the automated generation of specialized hardware modules. We will present the current version of the binary analysis and translation framework. Currently, our implementation is capable of producing ARMv8 and MicroBlaze (32-Bit) Executable and Linking Format (ELF) files or instruction traces. The framework can interpret the instructions for these two ISAs, and detect different types of instruction patterns. After detection, segments are converted into CDFG representations exposing the underlying Instruction Level Parallelism which we aim to exploit via automated hardware generation. On-going work is addressing the extraction of cyclical execution traces or static code blocks, more methods of hardware generation.
More information ...

UB07.4 RETINE: A PROGRAMMABLE 3D STACKED VISION CHIP ENABLING LOW LATENCY IMAGE ANALYSIS
Authors: Stéphane Chevobbé 1, Maria Lepecq 2 and Laurent Millet 2
1CEA LIST, FR; 2CEA-Leti, FR
Abstract
We have developed and fabricated a 3D stacked imager called RETINE composed with 2 layers based on the replication of a programmable 3D tile in a matrix manner providing a highly parallel programmable architecture. This tile is composed by a 16x16 BSI binned pixels array with associated readout and 16 column ADC on the first layer coupled to an efficient SIMD processor of 16 PE on the second layer. The prototype of RETINE achieves high video rates, from 5500 fps in binned mode to 340 fps in full resolution mode. It operates at 80 MHz with 720 mW power consumption leading to 85 GOPS/W power efficiency. To highlight the capabilities of the RETINE chip we have developed a demonstration platform with an electronic board embedding a RETINE chip that film rolling disks. Three scenario are available: high speed image capture, slow motion and composed image capture with parallel processing during acquisition.
More information ...

UB07.5 UWB ACKATCK: HIJACKING DEVICES IN UWB INDOOR POSITIONING SYSTEMS
Authors: Baptiste Pestourie, Vincent Berouille and Nicolas Fourty, Université Grenoble Alpes, FR
Abstract
Various radio-based Indoor Positioning Systems (IPS) have been proposed during the last decade as solutions to GPS inconsistency in indoor environments. Among the different radio technologies proposed for this purpose, 802.15.4 Ultra-Wideband (UWB) is by far the most performant, reaching up to 10 cm accuracy with 1000 Hz refresh rates. UWB is a popular technology for applications such as assets tracking in industrial environments or robots/drones indoor navigation. However, some security flaws in 802.15.4 standard expose UWB positioning to attacks. In this demonstration, we show how an attacker can exploit a vulnerability on 802.15.4 acknowledgment frames to hijack a device in a UWB positioning system. We demonstrate that using simply one cheap UWB chip, the attacker can take control over the positioning system and generate fake trajectories from a laptop. The results are observed in real-time in the 3D engine monitoring the positioning system.
More information ...

UB07.6 DESIGN AUTOMATION FOR EXTENDED BURST-MODE AUTOMATA IN WORKCRAFT
Authors: Alex Chan, Alex Yakovlev, Danil Sokolov and Victor Khomenko, Newcastle University, GB
Abstract
Asynchronous circuits are known to have high performance, robustness and low power consumption, which are particularly beneficial for the area of so-called "little digital" controllers where low latency is crucial. However, asynchronous design is not widely adopted by industry, partially due to the steep learning curve inherent in the complexity of formal specifications, such as Signal Transition Graphs (STGs). In this demo, we promote a class of the Finite State Machine (FSM) model called Extended Burst-Mode (XBM) automata as a practical way to specify many asynchronous circuits. The XBM specification has been automated in the Workcraft toolkit (https://workcraft.org) with elaborate support for state encoding, conditionals and "don't care" signals. Formal verification and logic synthesis of the XBM automata is implemented via conversion to the established STG model, reusing existing methods and CAD tools. Tool support for the XBM flow will be demonstrated using several case studies.
More information ...

UB07.7 DEEPESENSE-FPGA: FPGA ACCELERATION OF A MULTIMODAL NEURAL NETWORK
Authors: Mehdi Trabelsi Ajili and Yuko Hara-Azumi, Tokyo Institute of Technology, JP
Abstract
Currently, Internet of Things and Deep Learning (DL) are merging into one domain and creating outstanding technologies for various classification tasks. Such technologies require complex DL networks that are mainly targeting powerful platforms with rich computing resources like servers. Therefore, for resource-constrained embedded systems, new challenges of size, performance and power consumption have to be considered, particularly when edge devices handle multimodal data, i.e., different types of real-time sensing data (voice, video, text, etc.). Our ongoing project is focused on DeepSense, a multimodal DL framework combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to process time-series data, such as accelerometer and gyroscope to detect human activity. We aim at accelerating DeepSense by FPGA (Xilinx Zynq) in a hardware-software co-design manner. Our demo will show the latest achievements through latency and power consumption evaluations.
More information ...

UB07.8 SUBRISC+: IMPLEMENTATION AND EVALUATION OF AN EMBEDDED PROCESSOR FOR LIGHTWEIGHT IOT EHEALTH
Authors: Mingyu Yang and Yuko Hara-Azumi, Tokyo Institute of Technology, JP
Abstract
Although the rapid growth of Internet of Things (IoT) has enabled new opportunities for eHealth devices, the further development of complex systems is severely constrained by the power and energy supply on the battery-powered embedded systems. To address this issue, this work presents a processor design called "SubRISC+", targeting lightweight IoT eHealth. SubRISC+ is a processor design to achieve low power/energy consumption through its unique and compact architecture. As an example of lightweight eHealth applications on SubRISC+, we are working on the epileptic seizure detection using the dynamic time wrapping algorithm to deploy on wearable IoT eHealth devices. Simulation results show that 22% reduction on dynamic power and 50% reduction on leakage power and core area are achieved compared to Cortex-M0. As an ongoing work, the evaluation on a fabricated chip will be done within the first half of 2020.
More information ...
PA-HLS: HIGH-LEVEL ANNOTATION OF ROUTING CONGESTION FOR XILINX VIVADO HLS DESIGNS

Authors: Osama Bin Tariq1, Junnan Shan1, Luciano Lavagno1, Georgios Floros2, Mihai Teodor Lazarescu1, Christos Sotiriou1 and Mario Roberto Casu1
1Politecnico di Torino, IT; 2University of Thessaly, GR

Abstract
We will demo a novel high-level backannotation flow that reports routing congestion issues at the C++ source level by analyzing reports from FPGA physical design (Xilinx Vivado) and internal debugging files of the Vivado HLS tool. The flow annotates the C++ source code, identifying likely causes of congestion, e.g., on-chip memories or the DSP units. These shared resources often cause routing problems on FPGAs because they cannot be duplicated by physical design. We demonstrate on realistic large designs how the information provided by our flow can be used to both identify congestion issues at the C++ source level and solve them using HLS directives. The main demo steps are: 1- Extraction of the source-level debugging information from the Vivado HLS database 2- Generation of a list of net names involved in congestion areas and of their relative significance from the Vivado post-global-routing database 3- Visualization of the C++ code lines that contribute most to congestion.

More information ...

MDD-COP: A PRELIMINARY TOOL FOR MODEL-DRIVEN DEVELOPMENT EXTENDED WITH LAYER DIAGRAM FOR CONTEXT-ORIENTED PROGRAMMING

Authors: Harumi Watanabe1, Chintats Yamamoto1, Takeshi Ohkawa1, Mikiko Sato1, Nobuhiko Ogura1 and Mana Tabei1
1Tokai University, JP; 2Tokyo City University, JP

Abstract
This presentation introduces a preliminary tool for Model-Driven development (MDD) to generate programs for Context-Oriented Programming (COP). In modern embedded systems such as IoT and Industry 4.0, their software began to process multiple services by following the changing surrounding environments. COP is a powerful technique for developing such modern systems, the works of modeling for COP are limited. There are no works to mention the relation between UML (Unified Modeling Language) and COP. To solve this problem, we provide a COP generation from a layer diagram extended the package diagram of UML by stereotypes. In our approach, users draw a layer diagram and other UML diagrams, then xUML, which is a major tool of MDD, generates XML code with layer information for COP; finally, our tool generates COP code from XML code.

More information ...

EVENT-BASED AI FOR AUTOMOTIVE AND IOT

Speaker: Etienne Perot, Prophesee, FR
Author: Etienne Perot, Prophesee, FR

Abstract
Event cameras are a new type of sensor encoding visual information in the form of asynchronous events. An event corresponds to a change in the log-luminosity intensity at a given pixel location. Compared to standard frame cameras, event cameras have higher temporal resolution, higher dynamic range and lower power consumption. Thanks to these characteristics, event cameras find many applications in automotive and IoT, where low latency, robustness to challenging lighting conditions and low power consumption are critical requirements. In this talk we present recent advances in artificial intelligence applied to event cameras. In particular, we discuss how to adapt deep learning methods to work on events and their advantages compared to conventional frame-based methods. The presentation will be illustrated by results on object detection in automotive and IoT scenarios, running real-time on mobile platforms.
7.2 Reconfigurable Systems and Architectures

Date: Wednesday 11 March 2020
Time: 14:30 - 16:00
Location / Room: Chamrousse
Chair: Christian Pilato, Politecnico di Milano, IT
Co-Chair: Philippe Coussy, University Bretagne Sud / Lab-STI, FR

Reconfigurable technologies are evolving at the device, architecture, and system levels, from embedded computation to server-based accelerator integration. In this session we explore ideas at these levels, discussing architectural features for power optimisation of CGRAs, a framework for integrating FPGA accelerators in serverless environments, and placement strategies on alternative FPGA device technologies.

14:15 7.1.3 NEURONFLOW: A NEUROMORPHIC PROCESSOR ARCHITECTURE FOR LIVE AI APPLICATIONS
Speaker: Orlando Moreira, GraAL Matter Labs, NL
Authors: Orlando Moreira, Amirreza Yousefzadeh, Gokturk Cinerin, Rik-Jan Zwartenkot, Ajay Kapoor, Fabian Chersi, Peng Qiao, Peter Kevits, Mina Khoei, Louis Rouillard, Ashoka Visweswara and Jonathan Tapson, GraAL Matter Labs, NL
Abstract
This paper gives an overview of the Neuronflow many-core architecture. It is a neuromorphic data flow architecture that exploits brain-inspired concepts to deliver a scalable event-based processing engine for neuron networks in Live AI applications at the edge. Its design is inspired by brain biology, but not necessarily biologically plausible. The main design goal is the exploitation of sparsity to dramatically reduce latency and power consumption as required by sensor processing at the Edge.

Download Paper (PDF; Only available from the DATE venue WiFi)

14:36 7.1.4 SPECK - SUB-MW SMART VISION SENSOR FOR MOBILE IOT APPLICATIONS
Author: Ning Qiao, aiCTX, CH
Abstract
Speck is the first available neuromorphic smart vision sensor system-on-chip (SoC), which combines neuromorphic vision sensing and neuromorphic computation on a single die, for mW vision processing. The DVS pixel array is coupled directly to a new fully-asynchronous event-driven spiking CNN processor for highly compact and energy efficient dynamic visual processing. Speck supports a wide range of potential applications, spanning industrial and consumer-facing use cases.

16:00 End of session
ENERGY-AWARE PLACEMENT FOR SRAM-NVM HYBRID FPGAS

Speaker: Seongskik Park, Seoul National University, KR
Authors: Seongskik Park, Jongwan Kim and Sungroh Yoon, Seoul National University, KR

Abstract
Field-programmable gate arrays (FPGAs) have been widely used in many applications due to their reconfigurability. Especially, the short development time makes the FPGAs one of the promising reconfigurable architectures for emerging applications, such as deep learning. As CMOS technology advances, however, conventional SRAM-based FPGAs have approached their limitations. To overcome these obstacles, NVM-based FPGAs have been introduced. Although NVM-based FPGAs have the features of high area density, low static power consumption, and non-volatility, they are struggling to reduce energy consumption. Their challenge is mainly caused by the access speed of NVM, which is relatively slower than SRAM. In this paper, for compensating this limitation, we suggest SRAM-NVM hybrid FPGA architecture with SRAM- and NVM-based CLBs. In addition, we propose an energy-aware placement for efficient use of the SRAM-NVM hybrid FPGAs. As a result of our experiments, we were able to reduce the average energy consumption of SRAM-NVM hybrid FPGA by 22.23% and 21.94% compared to SRAM-based FPGA on the MCNC and VTR benchmarks, respectively.

Download Paper (PDF; Only available from the DATE venue WiFi)

REALIZING QUANTUM CIRCUITS ON IBM Q DEVICES

Speaker: Robert Wille, Johannes Kepler University Linz, AT
Authors: Carmen G. Almudever, TU Delft, NL

Abstract
A number of quantum computing devices consisting of a few tens of noisy qubits already exist. All of them present various limitations such as limited qubit connectivity and reduced gate set that must be considered to make quantum algorithms executable. In this talk, after briefly introduce the basics of quantum computing, we will provide an overview on the problem of realizing quantum circuits. We will discuss different mapping approaches as well as quantum devices emphasizing their main constraints. Special attention will be given to the quantum chips developed within the QuTech-Intel partnership.

EVERY DEVICE IS (ALMOST) EQUAL BEFORE THE COMPILER

Speaker: Gian Giacomo Guerreschi, Intel Corporation, US
Authors: Carmen G. Almudever, TU Delft, NL

Abstract
In 2017, IBM launched the first publicly available quantum computing device which is accessible through a cloud service. In the meantime, many further devices followed which have been used by more than 100,000 people who have executed more than 7 million experiments on them. Accordingly, fast and efficient solutions to realize quantum functionality to those devices are demanded by a huge user-base. This talk will provide an overview on IBM’s own tools for that matter as well solutions which have been developed by researchers world-wide - including a description of a compiler that won the IBM Qiskit Developer Challenge.

This session presents recent concerns and innovative solutions in verification and simulation, covering topics ranging from partial verification to lazy event prediction, till signal name disambiguation. They tackle these challenges by reducing complexity, exploiting GPUs, and using similarity-learning techniques.
**15:00** 7.4.2  **GPU-Accelerated Time Simulation of Systems with Adaptive Voltage and Frequency Scaling**

**Speaker:** Eric Schneider, University of Stuttgart, DE

**Authors:** Eric Schneider and Hans-Joachim Wunderlich, University of Stuttgart, DE

**Abstract**

Timing validation of systems with adaptive voltage- and frequency scaling (AVFS) requires an accurate timing model under multiple operating points. Simulating such a model at gate level is extremely time-consuming, and the state-of-the-art compromises both accuracy and compute efficiency. This paper presents a method for dynamic gate delay modeling on graphics processing units (GPU) accelerators which is based on polynomial approximation with offline statistical learning using regression analysis. It provides glitch-accurate switching activity information for gates and designs under varying supply voltages with negligible memory and performance impact. Parallelism from the evaluation of operating conditions, gates and stimuli is exploited simultaneously to utilize the high arithmetic computing throughput of GPUs. This way, large-scale design space exploration of AVFS-based systems is enabled. Experimental results demonstrate the efficiency and accuracy of the presented approach showing speedups of three orders of magnitude over conventional time simulation that supports static delays only.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

**15:30** 7.4.3  **Lazy Event Prediction Using Defining Trees and Schedule Bypass for Out-of-Order PDES**

**Speaker:** Rainer Doemer, University of California, Irvine, US

**Authors:** Daniel Mendoza, Zhongqi Cheng, Emad Arasteh and Rainer Doemer, University of California, Irvine, US

**Abstract**

Out-of-order parallel discrete event simulation (PDES) has been shown to be very effective in speeding up system design by utilizing parallel processors on multi- and many-core hosts. As the number of threads in the design model grows larger, however, the original scheduling approach does not scale. In this work, we analyze the out-of-order scheduler and identify a bottleneck with quadratic complexity in event prediction. We propose a more efficient lazy strategy based on defining trees and a schedule bypass with $O(m \log_2 m)$ complexity which shows sustained and improved performance gains in simulation of SystemC models with many processes. For models containing over 1000 processes, experimental results show simulation run time speedups of up to 90x using lazy event prediction against the original out-of-order PDES approach.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

**16:00** 7.4.4  **Towards Specification and Testing of RISC-V ISA Compliance**

**Speaker:** Vladimir Herdt, University of Bremen, DE

**Authors:** Vladimir Herdt¹, Daniel Grosse² and Rolf Drechsler ²

**Abstract**

Compliance testing for RISC-V is very important. Therefore, an official hand-written compliance test-suite is being actively developed. However, this requires significant manual effort in particular to achieve a high test coverage. In this paper we propose a test-suite specification mechanism in combination with a first set of instruction constraints and coverage requirements for the base RISC-V ISA. In addition, we present an automated method to generate a test-suite that satisfies the specification. Our evaluation demonstrates the effectiveness and potential of our method.

Download Paper (PDF; Only available from the DATE venue WiFi)
In the era of heterogeneous embedded systems, the diverse nature of computing elements pushes more than ever the need for smart runtime systems to be able to deal with resource management, multi-application mapping, task parallelism, and non-functional constraints. This session tackles these issues with solutions that span from resource-aware software architectures to novel runtime systems optimizing memory and energy consumption.

### Abstract

Due to the complexity of designs, post-silicon validation remains a major challenge with few systematic solutions. We provide an overview of the state-of-the-art post silicon validation process used by IBM to verify its latest IBM POWER9 processor. During the POWER9 post-silicon validation, we detected and handled 30% more logic bugs in 80% of the time, as compared to the previous IBM POWER8 bring-up. This improvement is the result of lessons learned from previous designs, leading to numerous innovations. We provide bug analysis data and compare it to POWER8 results. We demonstrate our methodology by describing several bugs from fail detection to root cause.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

### Authors

Robert Khasanov, TU Dresden, DE

---

### Title

**ENERGY-EFFICIENT RUNTIME RESOURCE MANAGEMENT FOR ADAPTABLE MULTI-APPLICATION MAPPING**

---

### Abstract

Modern embedded computing platforms consist of a high amount of heterogeneous resources, which allows executing multiple applications on a single device. The number of running application on the system varies with time and so does the amount of available resources. This has considerably increased the complexity of analysis and optimization algorithms for runtime mapping of firm real-time applications. To reduce the runtime overhead, researchers have proposed to pre-compute partial mappings at compile time and have the runtime efficiently compute the final mapping. However, most existing solutions only compute a fixed mapping for a given set of running applications, and the mapping is defined for the entire duration of the workload execution. In this work we apply algorithms to adapt to the amount of available resources by using mapping segments. This way, applications may switch between different configurations with varied degree of parallelism. We present a runtime manager for firm real-time applications that generates such mapping segments based on partial solutions and aims at minimizing the overall energy consumption without deadline violations. The proposed algorithm outperforms the state-of-the-art approaches on the overall energy consumption by up to 13% while incurring an order of magnitude less scheduling overhead.

Download Paper (PDF; Only available from the DATE venue WiFi)
Hardware architectures are under the continuous threat of all types of attacks. This session covers attacks based on side-channel leakage and the exploitation of vulnerabilities at the micro-architectural and circuit level.

**7.6 Attacks on Hardware Architectures**

**Date:** Wednesday 11 March 2020  
**Time:** 14:30 - 16:00  
**Location / Room:** Lesdiguières

**Chair:** Johanna Sepúlveda, Airbus Defence and Space, DE  
**Co-Chair:** Jean-Luc Danger, Télécom ParisTech, FR

Hardware architectures are under the continuous threat of all types of attacks. This session covers attacks based on side-channel leakage and the exploitation of vulnerabilities at the micro-architectural and circuit level.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
</table>
| 16:00      | IP3-119| **ON THE TASK MAPPING AND SCHEDULING FOR DAG-BASED EMBEDDED VISION APPLICATIONS ON HETEROGENEOUS MULTI/MANY-CORE ARCHITECTURES** | Nicola Bombieri, Università di Verona, IT  
**Authors:** Stefano Aldegheri¹, Nicola Bombieri¹ and Hiren Patel²  
¹Università di Verona, IT; ²University of Waterloo, CA |
|            |        | **Abstract**                                                                       |                                                                                               |
|            |        | In this work, we show that the applying the heterogeneous earliest finish time (HEFT) heuristic for the task scheduling of embedded vision applications can improve system performance up to 70% w.r.t. the scheduling solutions at the state of the art. We propose an algorithm called exclusive earliest finish time (XEFT) that introduces the notion of exclusion overlap between application primitives to improve the load balancing. We show that XEFT can improve the system performance up to 33% over HEFT, and 82% over the state of the art approaches. We present the results on different benchmarks, including a real-world localization and mapping application (ORB-SLAM) combined with the NVIDIA object detection application based on deep-learning. |                                                                                               |
|            |        | **Download Paper (PDF; Only available from the DATE venue WiFi)**                   |                                                                                               |
|            |        | 16:00                                **End of session**                              |                                                                                               |

**7.6.1 Sweeping for Leakage in Masked Circuit Layouts**

**Speaker:** Danilo Šijačić, IMEC / KU Leuven, BE  
**Authors:** Danilo Šijačić, Josep Balasch and Ingrid Verbauwhede, KU Leuven, BE

**Abstract**  
Masking schemes are the most popular countermeasure against side-channel analysis. They theoretically decorrelate information leaked through inherent physical channels from the key-dependent intermediate values that occur during computation. Their provable security is devised under models that abstract complex physical phenomena of the underlying hardware. In this work, we investigate the impact of the physical layout to the side-channel security of masking schemes. For this we propose a model for co-simulation of the analog power distribution network with the digital logic core. Our study considers the drive of the power supply buffers, as well as parasitic resistors, inductors and capacitors. We quantify our findings using Test Vector Leakage Assessment by relative comparison to the parasitic-free model. Thus we provide a deeper insight into the potential layout sources of leakage and their magnitude.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**7.6.2 Increased Reproducibility and Comparability of Data Leak Evaluations using ExOT**

**Speaker:** Philipp Medl, ETH Zürich, CH  
**Authors:** Philipp Medl¹, Bruno Klopott² and Lothar Thiele³  
¹ETH Zurich, CH; ²ETH Zürich, CH

**Abstract**  
As computing systems are increasingly shared among different users or application domains, researchers have intensified their efforts to detect possible data leaks. In particular, many investigations highlight the vulnerability of systems w. r. t. covert and side channel attacks. However, the effort required to reproduce and compare different results has proven to be high. Therefore, we present a novel methodology for covert channel evaluation. In addition, we introduce the Experiment Orchestration Toolkit ExOT, which provides software tools to efficiently execute the methodology. Our methodology ensures that the covert channel analysis yields expressive results that can be reproduced and allow the comparison of the threat potential of different data leaks. ExOT is a software bundle that consists of easy to extend C++ libraries and Python packages. These libraries and packages provide tools for the generation and execution of experiments, as well as the analysis of the experimental data. Therefore, ExOT decreases the engineering effort needed to execute our novel methodology. We verify these claims with an extensive evaluation of four different covert channels on an Intel Haswell and an ARMv8 based platform. In our evaluation, we derive capacity bounds and show achievable throughputs to compare the threat potential of these different covert channels.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

**7.6.3 Ghostbusters: Mitigating Spectre Attacks on a DBT-Based Processor**

**Speaker and Author:** Simon Rokicki, Irisa, FR

**Abstract**  
Unveiled early 2018, the Spectre vulnerability affects most of the modern high-performance processors. Spectre variants exploit the speculative execution mechanisms and a cache side-channel attack to leak secret data. As of today, the main countermeasures consist of turning off the speculation, which drastically reduces the processor performance. In this work, we focus on a different kind of micro-architecture: the DBT based processors, such as Transmeta Crusoe [1], Widaa Denver or Hybrid-DBT. Instead of using complex out-of-order (OoO) mechanisms, those cores combines a software Dynamic Binary Translation mechanism (DBT) and a parallel in-order architecture, typically a VLIW core. The DBT is in charge of translating and optimizing the binaries before their execution. Studies show that DBT based processors can reach the performance level of OoO cores for regular enough applications. In this paper, we demonstrate that, even if those processors do not use OoO execution, they are still vulnerable to Spectre variants, because of the DBT optimizations. However, we also demonstrate that those systems can easily be patched, as the DBT is done in software and has fine-grained control over the optimization process.

**Download Paper (PDF; Only available from the DATE venue WiFi)**
7.7 Self-Adaptive and Learning Systems

Date: Wednesday 11 March 2020
Time: 14:30 - 16:00
Location / Room: Berlioz

Chair: Gilles Sassatelli, Université de Montpellier, FR
Co-Chair: Rishad Shafik, University of Newcastle, GB

Recent advances in machine learning have pushed the boundaries of what is possible in self-adaptive and learning systems. This session pushes the state of art in runtime power and performance trade-offs for deep neural networks and self-optimizing embedded systems.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:30</td>
<td>7.6.4</td>
<td>DYNAMIC FAULTS BASED HARDWARE TROJAN DESIGN IN STT-MRAM</td>
<td>Sarath Mohanachandran Nair, Karlsruhe Institute of Technology, DE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker:</td>
<td>Sarath Mohanachandran Nair², Rajendra Bishnoi², Arunukumar Vijayan¹ and Mehdi Tahoori ¹</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Authors:</td>
<td>Karlsruhe Institute of Technology, DE; TU Delft, NL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abstract</td>
<td>The emerging Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM) is seen as a promising candidate to replace conventional on-chip memories. It has several advantages such as high density, non-volatility, scalability, and CMOS compatibility. With this technology becoming ubiquitous, it also becomes interesting as a target for security attacks. As the fabrication process of STT-MRAM evolves, it is susceptible to various fault mechanisms which are different from those of conventional CMOS memories. These unique fault mechanisms can be exploited by an adversary to deploy hardware Trojans, which are deliberately introduced design modifications. In this work, we demonstrate how a particular stealthy circuit modification to inject a fault mechanism, namely dynamic fault, can be exploited to implement a hardware Trojan trigger which cannot be detected by standard memory testing methods. The fault mechanisms can also be used to design new payloads specific to STT-MRAM. We illustrate this by proposing a new payload by utilizing coupling faults, which leads to degraded performance and data corruption. Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
</tbody>
</table>

| 15:45  | 7.6.5 | ORACLE-BASED LOGIC LOCKING ATTACKS: PROTECT THE ORACLE NOT ONLY THE NETLIST | Emmanouil Kalligeros, University of the Aegean, GR |
|        |       | Speaker:                                               | Emmanouil Kalligeros, Nikolaos Karousos and Irene Karybi, University of the Aegean, GR |
|        |       | Authors:                                               | Emmanouil Kalligeros, Nikolaos Karousos and Irene Karybi, University of the Aegean, GR |
|        |       | Abstract                                               | Logic locking has received a lot of attention in the literature due to its very attractive hardware-security characteristics: it can protect against IP piracy and overproduction throughout the whole IC supply chain. However, a large class of logic-locking attacks, the oracle-based ones, take advantage of a functional copy of the chip, the oracle, to extract the key that protects the chip. So far, the techniques dealing with oracle-based attacks focus on the netlist that the attacker possesses, assuming that the oracle is always available. For this reason, they are usually overcome by new attacks. In this paper, we propose a hardware security scheme that targets the protection of the oracle circuit, by locking the circuit when the necessary for setting the inputs and observing the outputs, scan in/out process begins. Hence, no correct input/output pairs can be acquired to perform the attacks. The proposed scheme is not based on controlling global signals like test_enable or scan_enable, whose values can be easily suppressed by the attacker. Security threats are identified, discussed and addressed. The developed scheme is combined with a traditional logic locking technique with high output corruptibility, to achieve increased levels of protection. Download Paper (PDF; Only available from the DATE venue WiFi) |

| 16:00  | IP3-  | ARE CLOUD FPGAS REALLY VULNERABLE TO POWER ANALYSIS ATTACKS? | Ognjen Glamocanin, EPFL, CH |
|        | 12,   | Speaker:                                               | Ognjen Glamocanin¹, Louis Coulon¹, Francesco Regazzoni² and Mirjana Stojilovic ¹ |
|        | 424   | Authors:                                               | EPFL, CH; ALaRI, CH |
|        |       | Abstract                                               | Recent works have demonstrated the possibility of extracting secrets from a cryptographic core running on an FPGA by means of remote power analysis attacks. To mount these attacks, an adversary implements a voltage fluctuation sensor in the FPGA logic, records the power consumption of the target cryptographic core, and recovers the secret key by running a power analysis attack on the recorded traces. Despite showing that the power analysis could also be performed without physical access to the cryptographic core, these works were mostly carried out on dedicated FPGA boards in a controlled environment, leaving open the question about the possibility to successfully mount these attacks on a real system deployed in the cloud. In this paper, we demonstrate, for the first time, a successful key recovery attack on an AES cryptographic accelerator running on an Amazon EC2 F1 instance. We collect the power traces using a delay-line based voltage drop sensor, adapted to the Xilinx Virtex Ultrascale+ architecture used on Amazon EC2 F1, where CARRY8 blocks do not have a monotonic delay increase at their outputs. Our results demonstrate that security concerns raised by multitenent FPGAs are indeed valid and that countermeasures should be put in place to mitigate them. Download Paper (PDF; Only available from the DATE venue WiFi) |

16:00   End of session
**15:00**  7.7.2 **ANTIDOTE: ATTENTION-BASED DYNAMIC OPTIMIZATION FOR NEURAL NETWORK RUNTIME EFFICIENCY**

**Speaker:** Xiang Chen, George Mason University, US  
**Authors:** Fuxun Yu¹, Chenchen Liu², Di Wang³, Yanzhi Wang¹ and Xiang Chen¹  
¹George Mason University, US; ²University of Maryland, Baltimore County, US; ³Microsoft, US

**Abstract**
Convolutional Neural Networks (CNNs) achieved great cognitive performance at the expense of considerable computation load. To relieve the computation load, many optimization works are developed to reduce the model redundancy by identifying and removing insignificant model components, such as weight sparsity and filter pruning. However, these works only evaluate model components’ static significance with internal parameter information, ignoring their dynamic interaction with external inputs. With per-input feature activation, the model component significance can dynamically change, and thus the static methods can only achieve sub-optimal results. Therefore, we propose a dynamic CNN optimization framework in this work. Based on the neural network attention mechanism, we propose a comprehensive dynamic optimization framework including (1) testing-phase channel and column feature map pruning, as well as (2) training-phase optimization by targeted dropout. Such a dynamic optimization framework has several benefits: (1) First, it can accurately identify and aggressively remove per-input feature redundancy with considering the model-input interaction; (2) Meanwhile, it can maximally remove the feature map redundancy in various dimensions thanks to the multi-dimension flexibility; (3) The training-testing co-optimization favors the dynamic pruning and helps maintain the model accuracy even with very high feature pruning ratio. Extensive experiments show that our method could bring 37.4%–44.5% FLOPs reduction with negligible accuracy drop on various of test networks.

Download Paper (PDF; Only available from the DATE venue WiFi)

**15:30**  7.7.3 **USING LEARNING CLASSIFIER SYSTEMS FOR THE DSE OF ADAPTIVE EMBEDDED SYSTEMS**

**Speaker:** Fedor Smirnov, Friedrich-Alexander-Universität Erlangen-Nürnberg, DE  
**Authors:** Fedor Smirnov, Behnaz Pourmohseni and Jürgen Teich, Friedrich-Alexander-Universität Erlangen-Nürnberg, DE

**Abstract**
Modern embedded systems are not only becoming more and more complex but are also often exposed to dynamically changing run-time conditions such as resource availability or processing power requirements. This trend has led to the emergence of adaptive systems which are designed using novel approaches that combine a static off-line Design Space Exploration (DSE) with the consideration of the dynamic run-time behavior of the system under design. In contrast to a static design approach, which provides a single design solution as a compromise between the possible run-time situations, the off-line DSE of these so-called hybrid design approaches yields a set of configuration alternatives, so that at run time, it becomes possible to dynamically choose the option most suited for the current situation. However, most of these approaches still use optimizers which were primarily developed for static design. Consequently, modern environment dynamic environments or run-time requirements is either not possible or comes at the cost of a significant computation overhead or results of poor quality. As a remedy, this paper introduces Learning Optimizer Constrained by Altering conditions (LOCAL), a novel optimization framework for the DSE of adaptive embedded systems. Following the structure of Learning Classifier System (LCS) optimizers, the proposed framework optimizes a strategy, i.e., a set of conditionally applicable solutions for the problem at hand, instead of a set of independent solutions. We show how the proposed framework—which can be used for the optimization of any adaptive system—is used for the optimization of dynamically reconfigurable many-core systems and provide experimental evidence that the hereby obtained strategy offers superior embeddability compared to the solutions provided by a s.o.t.a. hybrid approach which uses an evolutionary algorithm.

Download Paper (PDF; Only available from the DATE venue WiFi)

**16:00**  IP3-13  **EFFICIENT TRAINING ON EDGE DEVICES USING ONLINE QUANTIZATION**

**Speaker:** Michael Ostertag, University of California, San Diego, US  
**Authors:** Michael Ostertag¹, Sarah Al-Dowaisi² and Tajana Rosing¹  
¹University of California, San Diego, US; ²Kıng Abdulaziz City of Science and Technology, SA

**Abstract**
Sensor-specific calibration functions offer superior performance over global models and single-step calibration procedures but require prohibitive levels of sampling in the input feature space. Sensor self-calibration by gathering training data through collaborative calibration or self-analyzing predictive results allows these sensors to gather sufficient information. Resource-constrained edge devices are then stuck between high communication costs for transmitting training data to a centralized server and high memory requirements for storing data locally. We propose online dataset quantization that maximizes the diversity of input features, maintaining a representative set of data from a larger stream of training data points. We test the effectiveness of online dataset quantization on two real-world datasets: air quality calibration and power prediction modeling. Online Dataset Quantization outperforms reservoir sampling and performs equally to offline methods.

Download Paper (PDF; Only available from the DATE venue WiFi)
In this paper, we propose a novel one-hot representation for weights and activations in CNN model and demonstrate its benefits on hardware accelerator design. Specifically, rather than merely reducing the bitwidth, we quantize both weights and activations into n-bit integers that containing only one non-zero bit per value. In this way, the massive multiply and accumulates (MACs) are equivalent to additions of powers of two that can be efficiently calculated with histogram based computations. Experiments on the ImageNet classification task show that comparable accuracy can be obtained on our proposed One-Hot Networks (OHN) compared to conventional fixed-point networks. As case studies, we evaluate the efficacy of one-hot data representation on two state-of-the-art CNN accelerators on FPGA, our preliminary results show that 50% and 68.5% resource saving can be achieved on DaDianNao and Laconic respectively. Besides, the one-hot optimized Laconic can further achieve an average speedup of 4.94x on AlexNet.
In the past few years, Convolutional Neural Networks (CNNs) have seen a massive improvement, outperforming other visual recognition algorithms. Since they are playing an increasingly important role in fields such as face recognition, augmented reality or autonomous driving, there is the growing need for a fast and efficient system to perform the redundant and heavy computations of CNNs. This trend led researchers towards heterogeneous systems provided with hardware accelerators, such as GPUs and FPGAs. The vast majority of CNNs is implemented with floating-point parameters and operations, but from research, it has emerged that high classification accuracy can be obtained also by reducing the floating-point activations and weights to binary values. This context is well suitable for FPGAs, that are known to stand out in terms of performance when dealing with binary operations, as demonstrated in Finn, the state-of-the-art framework for building Binarized Neural Network (BNN) accelerators on FPGAs. In this paper, we propose a framework that extends Finn to a distributed scenario, enabling BNNs implementation on embedded multi-FPGA systems.

Download Paper (PDF; Only available from the DATE venue WiFi)
TOWARDS SPECIFICATION AND TESTING OF RISC-V ISA COMPLIANCE

Author: Vladimir Herdt, University of Bremen, DE

Abstract

Compliance testing for RISC-V is very important. Therefore, an official hand-written compliance test-suite is being actively developed. However, this requires significant manual effort in particular to achieve a high test coverage. In this paper we propose a test-suite specification mechanism in combination with a first set of instruction constraints and coverage requirements for the base RISC-V ISA. In addition, we present an automated method to generate a test-suite that satisfies the specification. Our evaluation demonstrates the effectiveness and potential of our method.

Download Paper (PDF; Only available from the DATE venue WiFi)

POST-SILICON VALIDATION OF THE IBM POWER9 PROCESSOR

Author: Hillel Mendelson, IBM, IL

Abstract

Due to the complexity of designs, post-silicon validation remains a major challenge with few systematic solutions. We provide an overview of the state-of-the-art post-silicon validation process used by IBM to verify its latest IBM POWER9 processor. During the POWER9 post-silicon validation, we detected and handled 30% more logic bugs in 80% of the time, as compared to the previous IBMPOWER8 bring-up. This improvement is the result of lessons learned from previous designs, leading to numerous innovations. We provide bug analysis data and compare it to POWER8 results. We demonstrate our methodology by describing several bugs from fail detection to root cause.

Download Paper (PDF; Only available from the DATE venue WiFi)

ON THE TASK MAPPING AND SCHEDULING FOR DAG-BASED EMBEDDED VISION APPLICATIONS ON HETEROGENEOUS MULTI/MANY-CORE ARCHITECTURES

Author: Nicola Bombieri, Università di Verona, IT

Abstract

In this work, we show that applying the heterogeneous earliest finish time (HEFT) heuristic for the task scheduling of embedded vision applications can improve the system performance up to 70% w.r.t. the scheduling solutions at the state of the art. We propose an algorithm called exclusive earliest finish time (XEFT) that introduces the notion of exclusive overlap between application primitives to improve the load balancing. We show that XEFT can improve the system performance up to 33% over HEFT, and 82% over the state of the art approaches. We present the results on different benchmarks, including a real-world localization and mapping application (ORB-SLAM) combined with the NVIDIA object detection application based on deep-learning.

Download Paper (PDF; Only available from the DATE venue WiFi)

ARE CLOUD FPGAS REALLY VULNERABLE TO POWER ANALYSIS ATTACKS?

Author: Ognjen Glamocanin, EPFL, CH

Abstract

Recent works have demonstrated the possibility of extracting secrets from a cryptographic core running on an FPGA by means of remote power analysis attacks. To mount these attacks, an adversary implements a voltage fluctuation sensor in the FPGA logic, records the power consumption of the target cryptographic core, and recovers the secret key by running a power analysis attack on the recorded traces. Despite showing that the power analysis could also be performed without physical access to the cryptographic core, these works were mostly carried out on dedicated FPGA boards in a controlled environment, leaving open the question about the possibility to successfully mount these attacks on a real system deployed in the cloud. In this paper, we demonstrate, for the first time, a successful key recovery attack on an AES cryptographic accelerator running on an Amazon EC2 F1 instance. We collect the power traces using a delay-line based voltage drop sensor, adapted to the Xilinx Virtex Ultrascale+ architecture used on Amazon EC2 F1, where CARRY8 blocks do not have a monotonic delay increase at their outputs. Our results demonstrate that security concerns raised by multitenant FPGAs are indeed valid and that countermeasures should be put in place to mitigate them.

Download Paper (PDF; Only available from the DATE venue WiFi)

EFFICIENT TRAINING ON EDGE DEVICES USING ONLINE QUANTIZATION

Author: Michael Ostertag, University of California, San Diego, US

Abstract

Sensor-specific calibration functions offer superior performance over global models and single-step calibration procedures but require prohibitive levels of sampling in the input feature space. Sensor self-calibration by gathering training data through collaborative calibration or self-analyzing predictive results allows these sensors to gather sufficient information. Resource-constrained edge devices are then stuck between high communication costs for transmitting training data to a centralized server and high memory requirements for storing data locally. We propose online dataset quantization that maximizes the diversity of input features, maintaining a representative set of data from a larger stream of training data points. We test the effectiveness of online dataset quantization on two real-world datasets: air quality calibration and power prediction modeling. Online Dataset Quantization outperforms reservoir sampling and performs equally to offline methods.

Download Paper (PDF; Only available from the DATE venue WiFi)
MULTI-AGENT ACTOR-CRITIC METHOD FOR JOINT DUTY-CYCLE AND TRANSMISSION POWER CONTROL

Speaker:
Sota Sawaguchi, CEA-Leti, FR

Authors:
Sota Sawaguchi1, Jean-Frédéric Christmann2, Anca Molinos2, Carolyn Bernier1 and Suzanne Lesecq2

1CEA, FR; 2CEA-Leti, FR

Abstract
Energy-harvesting wireless networks (EH-IoT) have gained attention due to their infinite operation and maintenance-free property. However, maintaining energy neutral operation (ENO) of EH-IoT devices, such that the harvested and consumed energy are matched during a certain time period, is crucial. Guaranteeing this ENO condition and optimal power-performance trade-off under various workloads and transient wireless channel quality is particularly challenging. This paper proposes a multi-agent actor-critic method for modulating both the transmission duty-cycle and the transmitter output power based on the state-of-buffer (SoB) and the state-of-charge (SoC) information as a state. Thanks to these buffers, system uncertainties, especially harvested energy and wireless link conditions, are addressed effectively. Differently from the state-of-the-art, our solution does not require any model of the wireless transceiver nor any measurement of wireless channel quality. Simulation results of a solar powered EH-IoT node using real-life outdoor solar irradiance data show that the proposed method achieves better performance without system failure throughout a year compared to the state-of-the-art that suffers some system downtime. Our approach also predicts almost no system failure during five years of operation. This proves that our approach can adapt to the change in energy-harvesting and wireless channel quality, all without direct observations.

Download Paper (PDF; Only available from the DATE venue WiFi)
The critical importance of testability for electronic devices led to the development of IEEE test standards. These methods, if not protected, offer a security backdoor to attackers. This demonstrator illustrates a state-of-the-art solution that prevents unauthorized usage of the test infrastructure based on the IEEE 1687 standard and implemented on an FPGA target.

More information ...

We have developed and fabricated a 3D stacked imager called RETINE composed with 2 layers based on the replication of a programmable 3D tile in a matrix providing a highly parallel programmable architecture. This tile is composed by a 16x16 BSI binned pixels array with associated readout and 16 column ADC on the first layer coupled to an efficient SiMMD processor of 16 PE on the second layer. The prototype of RETINE achieves high video rates, from 5500 fps in binned mode to 340 fps in full resolution mode. It operates at 80 MHz with 720 mW power consumption leading to 85 GOPS/W power efficiency. To highlight the capabilities of the RETINE chip we have developed a demonstration platform with an electronic board embedding a RETINE chip that films rotating disks. Three scenarios are available: high speed image capture, slow motion and composed image capture with parallel processing during acquisition.

More information ...

Recently, owing to the scaling down of technology and 2.5D/3D integration, power densities and temperatures of chips have been increasing significantly. Though commercial computational fluid dynamics tools can provide accurate thermal maps, they may lead to inefficiency in thermal-aware design with huge runtime. Thus, we develop the chip/package/system/level thermal analyzer, called FastThermSim, which can assist you to improve your design under thermal constraints in pre/post-silicon stages. In FastThermSim, we consider three heat transfer modes, conduction, convection, and thermal radiation. We convert them to temperature-independent terms by linearization methods and build a compact thermal model (CTM). By applying numerical methods to the CTM, the steady-state and transient thermal profiles can be solved efficiently without loss of accuracy. Finally, an easy-to-use thermal analysis tool is implemented for your design, which is flexible and compatible, with the graphic user interface.

More information ...

We will demo a novel high-level backannotation flow that reports routing congestion issues at the C++ source level by analyzing reports from FPGA physical design (Xilinx Vivado) and internal debugging files of the Vivado HLS tool. The flow annotates the C++ source code, identifying likely causes of congestion, e.g., on-chip memories or the DSP units. These shared resources often cause routing problems on FPGAs because they cannot be duplicated by physical design. We demonstrate our realistic large designs how the information provided by our flow can be used to both identify congestion issues at the C++ source level and solve them using HLS directives. The main demo steps are: 1- Extraction of the source-level debugging information from the Vivado HLS database 2- Generation of a list of net names involved in congestion areas and of their relative significance from the Vivado post-global-routings database 3- Visualization of the C++ code lines that contribute most to congestion

More information ...

This presentation introduces a preliminary tool for Model-Driven development (MDD) to generate programs for Context-Oriented Programming (COP). In modern embedded systems such as IoT and Industry 4.0, their software began to process multiple services by following the changing surrounding environments. COP is helpful for programming such software. In COP, we can consider the surrounding environments and multiple services as contexts and layers. Even though MDD is a powerful technique for developing such modern systems, the works of modeling for COP are limited. There are no works to mention the relation between UML (Unified Modeling Language) and COP. To solve this problem, we provide a COP generation from a layer diagram extended the package diagram of UML by stereotypes. In our approach, users draw a layer diagram and other UML diagrams, then xtUML, which is a major tool of MDD, generates XML code with layer stereotypes. In our approach, users draw a layer diagram and other UML diagrams, then xtUML, which is a major tool of MDD, generates XML code with layer stereotypes. In our approach, users draw a layer diagram and other UML diagrams, then xtUML, which is a major tool of MDD, generates XML code with layer stereotypes. In our approach, users draw a layer diagram and other UML diagrams, then xtUML, which is a major tool of MDD, generates XML code with layer stereotypes.

More information ...

8.1 Special Day on “Embedded AI”: Neuromorphic chips and systems

Date: Wednesday 11 March 2020
Time: 17:00 - 18:30
Location / Room: Amphithéâtre Jean Prouve
Chair: Wei Lu, University of Michigan, US
Co-Chair: Bernabe Linares-Barranco, CSIC, ES

Within the global field of AI, there is a subfield that focuses on exploiting neuroscience knowledge for artificial intelligent hardware systems. This is the neuromorphic engineering field. This session presents some examples of AI research focusing on this AI subfield.

Time | Label | Presentation Title | Authors
--- | --- | --- | ---
18:00 | | End of session | |

1

UB08.6 SRSN: SECURE RECONFIGURABLE TEST NETWORK
Authors: Vincent Reynaud1, Emanuele Valea2, Paolo Maistri1, Regis Leveugle1, Marie-Lise Flottes1, Sophie Dupuis2, Bruno Rouzeyle2 and Giorgio Di Natale1
1TIMA Laboratory, FR; 2LRMM, FR

Abstract
The critical importance of testability for electronic devices led to the development of IEEE test standards. These methods, if not protected, offer a security backdoor to attackers. This demonstrator illustrates a state-of-the-art solution that prevents unauthorized usage of the test infrastructure based on the IEEE 1687 standard and implemented on an FPGA target.

More information ...

UB08.7 RETINE: A PROGRAMMABLE 3D STACKED VISION CHIP ENABLING LOW LATENCY IMAGE ANALYSIS
Authors: Stéphane Chevobbe1, Maria Lepecq2 and Laurent Millet2
1CEA LIST, FR; 2CEA-Leti, FR

Abstract
We have developed and fabricated a 3D stacked imager called RETINE composed with 2 layers based on the replication of a programmable 3D tile in a matrix providing a highly parallel programmable architecture. This tile is composed by a 16x16 BSI binned pixels array with associated readout and 16 column ADC on the first layer coupled to an efficient SiMMD processor of 16 PE on the second layer. The prototype of RETINE achieves high video rates, from 5500 fps in binned mode to 340 fps in full resolution mode. It operates at 80 MHz with 720 mW power consumption leading to 85 GOPS/W power efficiency. To highlight the capabilities of the RETINE chip we have developed a demonstration platform with an electronic board embedding a RETINE chip that films rotating disks. Three scenarios are available: high speed image capture, slow motion and composed image capture with parallel processing during acquisition.

More information ...

UB08.8 FASTHERMSIM: FAST AND ACCURATE THERMAL SIMULATIONS FROM CHIPLETS TO SYSTEM
Authors: Yu-Min Lee, Chi-Wen Pan, Li-Rui Ho and Hong-Wen Chiou, National Chiao Tung University, TW

Abstract
Recently, owing to the scaling down of technology and 2.5D/3D integration, power densities and temperatures of chips have been increasing significantly. Though commercial computational fluid dynamics tools can provide accurate thermal maps, they may lead to inefficiency in thermal-aware design with huge runtime. Thus, we develop the chip/package/system/level thermal analyzer, called FastThermSim, which can assist you to improve your design under thermal constraints in pre/post-silicon stages. In FastThermSim, we consider three heat transfer modes, conduction, convection, and thermal radiation. We convert them to temperature-independent terms by linearization methods and build a compact thermal model (CTM). By applying numerical methods to the CTM, the steady-state and transient thermal profiles can be solved efficiently without loss of accuracy. Finally, an easy-to-use thermal analysis tool is implemented for your design, which is flexible and compatible, with the graphic user interface.

More information ...

UB08.9 PA-HLS: HIGH-LEVEL ANNOTATION OF ROUTING CONGESTION FOR XILINX VIVADO HLS DESIGNS
Authors: Osama Bin Tariq1, Junnan Shan1, Luciano Lavagno1, Georgios Floros1, Mihai Teodor Lazarescu1, Christos Sotiriou2 and Mario Roberto Casu2
1Politecnico di Torino, IT; 2University of Thessaly, GR

Abstract
We will demo a novel high-level backannotation flow that reports routing congestion issues at the C++ source level by analyzing reports from FPGA physical design (Xilinx Vivado) and internal debugging files of the Vivado HLS tool. The flow annotates the C++ source code, identifying likely causes of congestion, e.g., on-chip memories or the DSP units. These shared resources often cause routing problems on FPGAs because they cannot be duplicated by physical design. We demonstrate on realistic large designs how the information provided by our flow can be used to both identify congestion issues at the C++ source level and solve them using HLS directives. The main demo steps are: 1- Extraction of the source-level debugging information from the Vivado HLS database 2- Generation of a list of net names involved in congestion areas and of their relative significance from the Vivado post-global-routings database 3- Visualization of the C++ code lines that contribute most to congestion

More information ...

UB08.10 MDD-COP: A PRELIMINARY TOOL FOR MODEL-DRIVEN DEVELOPMENT EXTENDED WITH LAYER DIAGRAM FOR CONTEXT-ORIENTED PROGRAMMING
Authors: Harumi Watanabe1, Chiharu Yamamoto1, Takeshi Ohkawa1, Mikiko Ogura1 and Mana Tabei1
1Tokai University, JP; 2Tokyo City University, JP

Abstract
This presentation introduces a preliminary tool for Model-Driven development (MDD) to generate programs for Context-Oriented Programming (COP). In modern embedded systems such as IoT and Industry 4.0, their software began to process multiple services by following the changing surrounding environments. COP is helpful for programming such software. In COP, we can consider the surrounding environments and multiple services as contexts and layers. Even though MDD is a powerful technique for developing such modern systems, the works of modeling for COP are limited. There are no works to mention the relation between UML (Unified Modeling Language) and COP. To solve this problem, we provide a COP generation from a layer diagram extended the package diagram of UML by stereotypes. In our approach, users draw a layer diagram and other UML diagrams, then xtUML, which is a major tool of MDD, generates XML code with layer information for COP; finally, our tool generates COP code from XML code.

More information ...

18:00 End of session
Automated test generation for Trojan detection using delay-based side-channel analysis

Speaker: Prabhat Mishra, University of Florida, US
Authors: Yangdi Lyu and Prabhat Mishra, University of Florida, US

Abstract
Side-channel analysis is widely used for hardware Trojan detection in integrated circuits by analyzing various side-channel signatures, such as timing, power, and path delay. Existing delay-based side-channel analysis techniques have two major bottlenecks: (i) they are not suitable in detecting Trojans since the delay difference between the golden design and a Trojan inserted design is negligible, and (ii) they are not effective in creating robust delay signatures due to reliance on random and ATPG based test patterns. In this paper, we propose an efficient test generation technique to detect Trojans using delay-based side channel analysis. This paper makes two important contributions. (1) We propose an automated test generation algorithm to produce test patterns that are likely to activate trigger conditions, and drastically change critical paths. Compared to existing approaches where delay difference is solely based on extra gates from a small Trojan, the change of critical paths by our approach will lead to significant difference in path delay. (2) We propose a fast and efficient reordering technique to maximize the delay deviation between the golden design and Trojan inserted design. Experimental results demonstrate that our approach significantly outperforms state-of-the-art approaches that rely on ATPG or random test patterns for delay-based side-channel analysis.

Download Paper (PDF; Only available from the DATE venue WiFi)
8.2 Optimizing System-Level Design for Machine Learning

Date: Wednesday 11 March 2020
Time: 17:00 - 18:30
Location / Room: Autrans

Chair: Luciano Lavagno, Politecnico di Torino, IT
Co-Chair: Yuko Hara-Azumi, Tokyo Institute of Technology, JP

In the last years, the use of ML techniques, as deep neural networks, have become a trend in system-level design, either to help the flow finding promising solutions or to deploy ML-based applications. This session presents various approaches to optimize several aspects of system-level design, like the mapping of applications on heterogeneous platforms, the inference of CNNs or the file-system usage.
Nowadays, SSD cache plays an important role in cloud storage systems. The associated write policy, which enforces an admission control policy regarding fillup data into the cache, has a significant impact on the performance of the cache system and the amount of write traffic to SSD caches. Based on our analysis on a typical cloud block storage system, approximately 47.09% writes are write-only, i.e., writes to the blocks which have not been read during a certain time window. Naively writing the write-only data to the SSD cache unnecessarily introduces a large number of harmful writes to the SSD cache without any contribution to cache performance. On the other hand, it is a challenging task to identify and filter out those write-only data in a real-time manner, especially in a cloud environment running changing and diverse workloads. In this paper, to alleviate the above cache problem, we propose an ML-WP, Machine Learning Based Write Policy, which reduces write traffic to SSDs by avoiding writing write-only data. The main challenge in this approach is to identify write-only data in a real-time manner. To realize ML-WP and achieve accurate write-only data identification, we use machine learning methods to identify write-only data in a real-time manner. To address this issue, we propose probabilistic SMBO, which selects candidates based on probabilistic estimation of their Pareto efficiency. With a formulation that incorporates error in accuracy prediction and uncertainty in latency measurement, probabilistic Pareto efficiency quantifies a candidate’s quality in two ways: its likelihood of being Pareto optimal, and the expected number of current Pareto optimal solutions that it will dominate. We evaluate our proposed method on four image classification problems. Compared to a deterministic approach, probabilistic SMBO consistently generates Pareto optimal solutions that perform better, and that are competitive with state-of-the-art efficient CNN models, offering tremendous speedup in inference latency while maintaining comparable accuracy.

We observe that the running time of fragmented files is 2.36X longer than that of continuous files and that F2FS's in-place update scheme is incapable of reducing fragmentation. A fragmented file system leads to a poor user experience. Reserving space to prevent fragmentation is an intuitive approach. However, reserving space for all files wastes space since there are a large number of files. To deal with this dilemma, we propose an adaptive reserved space (ARS) scheme to choose some specific files to update in the reserved space. How to effectively select reserved files is critical to performance. We collect file characteristics associated with fragmentation to construct data sets and use decision trees to accurately pick reserved files. Besides, adjustable reserved space and dynamic reservation strategy are adopted. We implement ARS on a HiKey960 development platform and a commercial smartphone with slight space and file creation time overheads. Experimental results show that ARS reduces file and free space fragmentation dramatically, improves file I/O performance and reduces garbage collection overhead compared to traditional F2FS and F2FS with in-place updates. Furthermore, ARS delivers up to 1.26X transactions per second under SQLite than traditional F2FS and reduces up to 41.72% running time of Facebook and Twitter than F2FS with in-place updates.

We evaluate our proposed method on four image classification problems. Compared to a deterministic approach, probabilistic SMBO consistently generates Pareto optimal solutions that perform better, and that are competitive with state-of-the-art efficient CNN models, offering tremendous speedup in inference latency while maintaining comparable accuracy.

We observe that the running time of fragmented files is 2.36X longer than that of continuous files and that F2FS's in-place update scheme is incapable of reducing fragmentation. A fragmented file system leads to a poor user experience. Reserving space to prevent fragmentation is an intuitive approach. However, reserving space for all files wastes space since there are a large number of files. To deal with this dilemma, we propose an adaptive reserved space (ARS) scheme to choose some specific files to update in the reserved space. How to effectively select reserved files is critical to performance. We collect file characteristics associated with fragmentation to construct data sets and use decision trees to accurately pick reserved files. Besides, adjustable reserved space and dynamic reservation strategy are adopted. We implement ARS on a HiKey960 development platform and a commercial smartphone with slight space and file creation time overheads. Experimental results show that ARS reduces file and free space fragmentation dramatically, improves file I/O performance and reduces garbage collection overhead compared to traditional F2FS and F2FS with in-place updates. Furthermore, ARS delivers up to 1.26X transactions per second under SQLite than traditional F2FS and reduces up to 41.72% running time of Facebook and Twitter than F2FS with in-place updates.

We evaluate our proposed method on four image classification problems. Compared to a deterministic approach, probabilistic SMBO consistently generates Pareto optimal solutions that perform better, and that are competitive with state-of-the-art efficient CNN models, offering tremendous speedup in inference latency while maintaining comparable accuracy.
8.4 Architectural and Circuit Techniques toward Energy-efficient Computing

**Presentation Title:** YOU ONLY SEARCH ONCE: A FAST AUTOMATION FRAMEWORK FOR SINGLE-STAGE DNN/ACCELERATOR CO-DESIGN  
**Speaker:** Weiwei Chen, Chinese Academy of Sciences, CN  
**Authors:** Weiwei Chen, Ying Wang, Shuang Yang, Cheng Liu and Lei Zhang, Chinese Academy of Sciences, CN  
**Abstract:** DNN/Accelerator co-design has shown great potential in improving QoR and performance. Typical approaches separate the design flow into two-stage: (1) designing an application-specific DNN model with high accuracy; (2) building an accelerator considering the DNN specific characteristics. However, it may fail in promising the highest composite score which combines the goals of accuracy and other hardware-related constraints (e.g., latency, energy efficiency) when building a specific neural-network-based system. In this work, we present a single-stage automated framework, YOSO, aiming to generate the optimal solution of software-and-hardware that flexibly balances between the goal of accuracy, power, and QoC. Compared with the two-stage method on the baseline systolic array accelerator and Cifar10 dataset, we achieve 1.42x~2.29x energy or 1.79x~3.07x latency reduction at the same level of precision, for different user-specified energy and latency optimization constraints, respectively.

**Presentation Title:** TRANSPIRE: AN ENERGY-EFFICIENT TRANSPRECISION FLOATING-POINT PROGRAMMABLE ARCHITECTURE  
**Speaker:** Rohit Prasad, Lab-SICC, UBS, France & DEI, UniBo, Italy, FR  
**Authors:** Rohit Prasad1, Satyajit Das2, Kevin Martin3, Giuseppe Tagliavini4, Philippe Cousseuy5, Luca Benini6 and Davide Rossi4  
1Université Bretagne Sud, FR; 2IIT Palakkad, IN; 3University Bretagne Sud, FR; 4Università di Bologna, IT; 5Université Bretagne Sud / Lab-STICC, FR; 6Università di Bologna and ETH Zurich, IT  
**Abstract:** In recent years, Coarse Grain Reconfigurable Architecture (CGRA) accelerators have been increasingly deployed in Internet-of-Things (IoT) end nodes. A modern CGRA has to support and efficiently accelerate both integer and floating-point (FP) operations. In this paper, we propose an ultra-low-power tunable-precision CGRA architectural template, called TRANSprecision floating-point Programmable architecture (TRANSPIRE), and its associated compilation flow supporting both integer and FP operations. TRANSPIRE employs transprecision computing and multiple Single Instruction Multiple Data (SIMD) to accelerate FP operations while boosting energy efficiency as well. Experimental results show that TRANSPIRE achieves a maximum of 10.06x performance gain and consumes 12.91x less energy w.r.t. a RISC-V based CPU with an enhanced ISA supporting SIMD-style vectorization and FP data-types, while executing applications for near-sensor computing and embedded machine learning, with an area overhead of 1.25x only.

**Presentation Title:** YOU ONLY SEARCH ONCE: A FAST AUTOMATION FRAMEWORK FOR SINGLE-STAGE DNN/ACCELERATOR CO-DESIGN  
**Speaker:** Weiwei Chen, Chinese Academy of Sciences, CN  
**Authors:** Weiwei Chen, Ying Wang, Shuang Yang, Cheng Liu and Lei Zhang, Chinese Academy of Sciences, CN  
**Abstract:** DNN/Accelerator co-design has shown great potential in improving QoR and performance. Typical approaches separate the design flow into two-stage: (1) designing an application-specific DNN model with high accuracy; (2) building an accelerator considering the DNN specific characteristics. However, it may fail in promising the highest composite score which combines the goals of accuracy and other hardware-related constraints (e.g., latency, energy efficiency) when building a specific neural-network-based system. In this work, we present a single-stage automated framework, YOSO, aiming to generate the optimal solution of software-and-hardware that flexibly balances between the goal of accuracy, power, and QoC. Compared with the two-stage method on the baseline systolic array accelerator and Cifar10 dataset, we achieve 1.42x~2.29x energy or 1.79x~3.07x latency reduction at the same level of precision, for different user-specified energy and latency optimization constraints, respectively.

**Presentation Title:** TRANSPIRE: AN ENERGY-EFFICIENT TRANSPRECISION FLOATING-POINT PROGRAMMABLE ARCHITECTURE  
**Speaker:** Rohit Prasad, Lab-SICC, UBS, France & DEI, UniBo, Italy, FR  
**Authors:** Rohit Prasad1, Satyajit Das2, Kevin Martin3, Giuseppe Tagliavini4, Philippe Cousseuy5, Luca Benini6 and Davide Rossi4  
1Université Bretagne Sud, FR; 2IIT Palakkad, IN; 3University Bretagne Sud, FR; 4Università di Bologna, IT; 5Université Bretagne Sud / Lab-STICC, FR; 6Università di Bologna and ETH Zurich, IT  
**Abstract:** In recent years, Coarse Grain Reconfigurable Architecture (CGRA) accelerators have been increasingly deployed in Internet-of-Things (IoT) end nodes. A modern CGRA has to support and efficiently accelerate both integer and floating-point (FP) operations. In this paper, we propose an ultra-low-power tunable-precision CGRA architectural template, called TRANSprecision floating-point Programmable architecture (TRANSPIRE), and its associated compilation flow supporting both integer and FP operations. TRANSPIRE employs transprecision computing and multiple Single Instruction Multiple Data (SIMD) to accelerate FP operations while boosting energy efficiency as well. Experimental results show that TRANSPIRE achieves a maximum of 10.06x performance gain and consumes 12.91x less energy w.r.t. a RISC-V based CPU with an enhanced ISA supporting SIMD-style vectorization and FP data-types, while executing applications for near-sensor computing and embedded machine learning, with an area overhead of 1.25x only.

**Presentation Title:** TRANSPIRE: AN ENERGY-EFFICIENT TRANSPRECISION FLOATING-POINT PROGRAMMABLE ARCHITECTURE  
**Speaker:** Rohit Prasad, Lab-SICC, UBS, France & DEI, UniBo, Italy, FR  
**Authors:** Rohit Prasad1, Satyajit Das2, Kevin Martin3, Giuseppe Tagliavini4, Philippe Cousseuy5, Luca Benini6 and Davide Rossi4  
1Université Bretagne Sud, FR; 2IIT Palakkad, IN; 3University Bretagne Sud, FR; 4Università di Bologna, IT; 5Université Bretagne Sud / Lab-STICC, FR; 6Università di Bologna and ETH Zurich, IT  
**Abstract:** In recent years, Coarse Grain Reconfigurable Architecture (CGRA) accelerators have been increasingly deployed in Internet-of-Things (IoT) end nodes. A modern CGRA has to support and efficiently accelerate both integer and floating-point (FP) operations. In this paper, we propose an ultra-low-power tunable-precision CGRA architectural template, called TRANSprecision floating-point Programmable architecture (TRANSPIRE), and its associated compilation flow supporting both integer and FP operations. TRANSPIRE employs transprecision computing and multiple Single Instruction Multiple Data (SIMD) to accelerate FP operations while boosting energy efficiency as well. Experimental results show that TRANSPIRE achieves a maximum of 10.06x performance gain and consumes 12.91x less energy w.r.t. a RISC-V based CPU with an enhanced ISA supporting SIMD-style vectorization and FP data-types, while executing applications for near-sensor computing and embedded machine learning, with an area overhead of 1.25x only.

**Presentation Title:** TRANSPIRE: AN ENERGY-EFFICIENT TRANSPRECISION FLOATING-POINT PROGRAMMABLE ARCHITECTURE  
**Speaker:** Rohit Prasad, Lab-SICC, UBS, France & DEI, UniBo, Italy, FR  
**Authors:** Rohit Prasad1, Satyajit Das2, Kevin Martin3, Giuseppe Tagliavini4, Philippe Cousseuy5, Luca Benini6 and Davide Rossi4  
1Université Bretagne Sud, FR; 2IIT Palakkad, IN; 3University Bretagne Sud, FR; 4Università di Bologna, IT; 5Université Bretagne Sud / Lab-STICC, FR; 6Università di Bologna and ETH Zurich, IT  
**Abstract:** In recent years, Coarse Grain Reconfigurable Architecture (CGRA) accelerators have been increasingly deployed in Internet-of-Things (IoT) end nodes. A modern CGRA has to support and efficiently accelerate both integer and floating-point (FP) operations. In this paper, we propose an ultra-low-power tunable-precision CGRA architectural template, called TRANSprecision floating-point Programmable architecture (TRANSPIRE), and its associated compilation flow supporting both integer and FP operations. TRANSPIRE employs transprecision computing and multiple Single Instruction Multiple Data (SIMD) to accelerate FP operations while boosting energy efficiency as well. Experimental results show that TRANSPIRE achieves a maximum of 10.06x performance gain and consumes 12.91x less energy w.r.t. a RISC-V based CPU with an enhanced ISA supporting SIMD-style vectorization and FP data-types, while executing applications for near-sensor computing and embedded machine learning, with an area overhead of 1.25x only.

**Presentation Title:** TRANSPIRE: AN ENERGY-EFFICIENT TRANSPRECISION FLOATING-POINT PROGRAMMABLE ARCHITECTURE  
**Speaker:** Rohit Prasad, Lab-SICC, UBS, France & DEI, UniBo, Italy, FR  
**Authors:** Rohit Prasad1, Satyajit Das2, Kevin Martin3, Giuseppe Tagliavini4, Philippe Cousseuy5, Luca Benini6 and Davide Rossi4  
1Université Bretagne Sud, FR; 2IIT Palakkad, IN; 3University Bretagne Sud, FR; 4Università di Bologna, IT; 5Université Bretagne Sud / Lab-STICC, FR; 6Università di Bologna and ETH Zurich, IT  
**Abstract:** In recent years, Coarse Grain Reconfigurable Architecture (CGRA) accelerators have been increasingly deployed in Internet-of-Things (IoT) end nodes. A modern CGRA has to support and efficiently accelerate both integer and floating-point (FP) operations. In this paper, we propose an ultra-low-power tunable-precision CGRA architectural template, called TRANSprecision floating-point Programmable architecture (TRANSPIRE), and its associated compilation flow supporting both integer and FP operations. TRANSPIRE employs transprecision computing and multiple Single Instruction Multiple Data (SIMD) to accelerate FP operations while boosting energy efficiency as well. Experimental results show that TRANSPIRE achieves a maximum of 10.06x performance gain and consumes 12.91x less energy w.r.t. a RISC-V based CPU with an enhanced ISA supporting SIMD-style vectorization and FP data-types, while executing applications for near-sensor computing and embedded machine learning, with an area overhead of 1.25x only.
8.5 CNN Dataflow Optimizations

Date: Wednesday 11 March 2020
Time: 17:00 - 18:30
Location / Room: Bayard

Chair: Mario Casu, Politecnico di Torino, IT
Co-Chair: Wanli Chang, University of York, GB

This session focuses on efficient dataflow approaches for reducing CNN runtime on embedded hardware platforms. The papers to be presented demonstrate techniques for enhancing parallelism to improve performance of CNNs, leverage output prediction to reduce the runtime for time-critical embedded applications during inference, and presents a Keras-based DNN framework for real-time cyber physical systems.

Time Label Presentation Title Authors
17:00 8.5.1 ANALYSIS AND SOLUTION OF CNN ACCURACY REDUCTION OVER CHANNEL LOOP TILING
Speaker: Yesung Kang, Pohang University of Science and Technology, KR
Authors: Yesung Kang1, Yoonho Park1, Sunghoon Kim1, Eunji Kwon1, Taeho Lim2, Mingyu Woo3, Sangyoon Oh4 and Seokhyeong Kang
1Pohang University of Science and Technology, KR; 2SK Hynix, KR; 3University of California, San Diego, US; 4UNIST, KR
Abstract
Owing to the growth of the size of convolutional neural networks (CNNs), quantization and loop tiling (also called loop breaking) are mandatory to implement CNN on an embedded system. However, channel loop tiling of quantized CNNs induces unexpected errors. We explain why channel loop tiling of quantized CNNs induces the unexpected errors, and how the errors affect the accuracy of state-of-the-art CNNs. We also propose a method to recover accuracy under channel tiling by compressing and decompressing the most-significant bits of partial sums. Using the proposed method, we can recover accuracy by 12.3% with only 1% circuit area overhead and an additional 2% of power consumption.
Download Paper (PDF; Only available from the DATE venue WiFi)

Download Paper (PDF; Only available from the DATE venue WiFi)
With the excellent accuracy and feasibility, Convolutional Neural Networks (CNNs) have been widely applied into novel intelligent applications and systems. However, the CNN computation performance is significantly hindered by its computation flow, which computes the model structure sequentially by layers with massive convolution operations. Such a layer-wise sequential computation flow is defined by the inter-layer data dependency and causes certain performance issues, such as resource under-utilization, significant computation overhead, etc. To solve these problems, in this work, we propose a novel CNN structural decoupling method, which could decouple CNN models by “critical paths” and eliminate the inter-layer data dependency. Based on this method, we redefine the CNN computation flow into parallel and cascade computing paradigms, which can significantly enhance the CNN computation performance with both multi-core and single-core CPU processors. Experiments show that our DC-CNN framework could reduce at most 33% latency on multi-core CPUs for both CIFAR and ImageNet. On small-capacity mobile platforms, cascade computing could reduce the latency by average 24% on ImageNet and 42% on CIFAR10. Meanwhile, the memory reduction could reach average 21% and 64%, respectively.

Download Paper (PDF; Only available from the DATE venue WiFi)
8.6 Microarchitecture-level reliability analysis and protection

Date: Wednesday 11 March 2020

Time: 17:00 - 18:30

Location / Room: Lesdiguières

Chair: Michail Maniatakos, New York University Abu Dhabi, UA

Co-Chair: Alessandro Savino, Politecnico di Torino, IT

Reliability analysis and protection at the microarchitecture level is of paramount importance to speed-up the design face of any computing system. On the analysis side, this session presents a reverse-order ACE (Architecturally Correct Execution) analysis that is more accurate than original ACE proposals, then moving to an instruction level analysis based on a genetic-algorithm able to improve program resiliency to errors. Finally, on the protection side, the session presents a low-cost ECC plus approximation mechanism for GPU register files.

17:00 8.6.1 RACE: REVERSE-ORDER PROCESSOR RELIABILITY ANALYSIS

Authors: Athanasios Chatzidimitriou and Dimitris Gizopoulos, University of Athens, GR

Abstract: Modern microprocessors suffer from increased error rates that come along with fabrication technology scaling. Processor designs continuously become more prone to hardware faults that lead to execution errors and system failures, which raise the requirement of protection mechanisms. However, error mitigation strategies have to be applied diligently, as they impose significant power, area, and performance overheads. Early and accurate reliability estimation of a microprocessor design is essential in order to determine the most vulnerable hardware structures and the most efficient protection schemes. One of the most commonly used techniques for reliability estimation is Architecturally Correct Execution (ACE) analysis. ACE analysis can be applied at different abstraction models, including microarchitecture and RTL and often requires a single or few simulations to report the Architectural Vulnerability Factor (AVF) of the processor structures. However, ACE analysis overestimates the vulnerability of structures because of its pessimistic, worst-case nature. Moreover, it only delivers coarse-grain vulnerability reports and no details about the expected result of hardware faults (silent data corruptions, crashes). In this paper, we present reverse ACE (rACE), a methodology that (a) improves the accuracy of ACE analysis and (b) delivers fine-grain error outcome reports. Using a reverse-order tracing flow, rACE analysis associates portions of the simulated execution of a program with the actual output and the control flow, delivering finer accuracy and results classification. Our findings show that rACE reports an average 1.45X overestimation, compared to Statistical Fault Injection, for reverse-order tracing flow, rACE analysis associates portions of the simulated execution of a program with the actual output and the control flow, delivering finer accuracy and results classification. Our findings show that rACE reports an average 1.45X overestimation, compared to Statistical Fault Injection, for different sizes of the register file of an out-of-order CPU core (executing both ARM and x86 binaries), when a baseline ACE analysis reports 2.3X overestimation and even refined versions of ACE analysis report an average of 1.8X overestimation.

Download Paper (PDF; Only available from the DATE venue WiFi)

17:30 8.6.2 DEFCON: GENERATING AND DETECTING FAILURE-PRONE INSTRUCTION SEQUENCES VIA STOCHASTIC SEARCH

Authors: Ioannis Tsokanos, Queen’s University Belfast, GB

Abstract: The increased variability and adopted low supply voltages render nanometer devices prone to timing failures, which threaten the functionality of digital circuits. Recent schemes focused on developing instruction-aware failure prediction models and adapting voltage/frequency to avoid errors while saving energy. However, such schemes may be inaccurate when applied to pipelined cores since they consider only the currently executed instruction and the preceding one, thereby neglecting the impact of all the concurrently executing instructions on failure occurrence. In this paper, we first demonstrate that the order and type of instructions in sequences with a length equal to the pipeline depth affect significantly the failure rate. To overcome the practically impossible evaluation of the impact of all possible sequences on failures, we present DEFCON, a fully automated framework that stochastically searches for the most failure-prone instruction sequences (ISQs). DEFCON generates such sequences by integrating a properly formulated genetic algorithm with accurate post-layout dynamic timing analysis, considering the data-dependent path sensitization and instruction execution history. The generated microarchitecture-aware ISQs are then used by DEFCON to estimate the failure vulnerability of any application. To evaluate the efficacy of the proposed framework, we implement a pipelined floating-point unit and perform dynamic timing analysis based on input data that we extract from a variety of applications consisting of up-to 43.5M ISQs. Our results show that DEFCON reveals quickly ISQs that maximize the output quality loss and correctly detects 99.7% of the actual faulty ISQs in different applications under various levels of variation-induced delay increase. Finally, DEFCON enables us to identify failure-prone ISQs early at the design cycle, and save 26.8% of energy on average when combined with a clock stretching mechanism.

Download Paper (PDF; Only available from the DATE venue WiFi)
In this paper, we present a serial-equivalent parallel router for FPGAs on modern multi-core processors. We are based on the inherent net order of serial router. Notably, our parallel router generates exactly the same wirelength as the serial router satisfying serial equivalency. Experimental results show that our parallel router provides about 19.13x speedup on average using 32 processor cores comparing to the serial router. Synchronization of conflicting stages using MPI-based message queue for a feasible routing solution. Note that load balance is always used to guide the router to schedule all the nets into a series of stages, where the non-conflicting nets are scheduled in the same stage and the conflicting nets are scheduled in different stages. We explore the parallel routing of non-conflicting nets on multi-core processors for a significant speedup. We perform the data synchronization of conflicting stages using MPI-based message queue for a feasible routing solution. Note that load balance is always used to guide the multi-core parallel routing. Experimental results show that our parallel router provides about 19.13x speedup on average using 32 processor cores compared to the serial router. Notably, our parallel router generates exactly the same wirelength as the serial router satisfying serial equivalency.
9.1 Special Day on “Silicon Photonics”: Advancements on Silicon Photonics

**Date:** Thursday 12 March 2020  
**Time:** 08:30 - 10:00  
**Location / Room:** Amphithéâtre Jean Prouve

**Chair:**  
Gabriela Nicolescu, Polytechnique Montréal, CA

**Co-Chair:**  
Luca Ramini, Hewlett Packard Labs, US

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
<th>Download Paper (PDF; Only available from the DATE venue WiFi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18:00</td>
<td>8.7.3</td>
<td><strong>SELF-ALIGNED DOUBLE-PATTERNING AWARE LEGALIZATION</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Author:** Hua Xiang, IBM Research, US  
**Authors:** Hua Xiang,1, Gi-Joon Nam,1, Gustavo Tellez,2, Shyam Ramji,2 and Xiaoping Xu,3  
1IBM Research, US; 2IBM Thomas J. Watson Research Center, US; 3University of Texas at Austin, US  
**Abstract**  
Double patterning is a widely used technique for sub-22nm. Among various double patterning techniques, Self-Aligned Double Patterning (SADP) is a promising technique for good mask overlay control. Based on SADP, a new set of standard cells (T-cells) are developed using thicker metal wires for stronger drive strength. By applying this kind of gates on critical paths, it helps to improve the design performance. However, a mixed design with T-cells and normal cells (N-cells) requires that T-cells are placed on circuit rows with thicker metal, and the normal cells are on the normal circuit rows. Therefore, a placer is needed to adjust the cells to the matched circuit rows. In this paper, a two-stage min-cost max-flow based legalization flow is presented to adjust N/T gate locations for a legal placement. The experimental results demonstrate the effectiveness and efficiency of our approach. | Download Paper (PDF; Only available from the DATE venue WiFi) |

| 18:15 | 8.7.4 | **EXPLAINABLE DRC HOTSPOT PREDICTION WITH RANDOM FOREST AND SHAP TREE EXPLAINER** |  
**Speaker:** Wei Zeng, University of Wisconsin-Madison, US  
**Authors:** Wei Zeng,1, Azadeh Davoodi,1 and Rasit Onur Topaloglu,2  
1University of Wisconsin - Madison, US; 2IBM, US  
**Abstract**  
With advanced technology nodes, resolving design rule check (DRC) violations has become a cumbersome task, which makes it desirable to make predictions at earlier stages of the design flow. In this paper, we show that the Random Forest (RF) model is quite effective for the DRC hotspot prediction at the global routing stage, and in fact significantly outperforms recent prior works, with only a fraction of the runtime to develop the model. We also propose, for the first time, to adopt a recent explanatory metric—the SHAP value—to make accurate and consistent explanations for individual DRC hotspot predictions from RF. Experiments show that RF is 21%-60% better in predictive performance on average, compared with promising machine learning models used in similar works (e.g. SVM and neural networks) while exhibiting good explainability, which makes it ideal for DRC hotspot prediction. | Download Paper (PDF; Only available from the DATE venue WiFi) |

| 18:31 | IP4-10 | **XGBIR: AN XBOOST-BASED IR DROP PREDICTOR FOR POWER DELIVERY NETWORK** |  
**Speaker:** Chi-Hsien Pao, Yu-Min Lee and An-Yu Su, National Chiao Tung University, TW  
**Authors:** Wei Zeng,1, Azadeh Davoodi,1 and Rasit Onur Topaloglu,2  
1University of Wisconsin - Madison, US; 2IBM, US  
**Abstract**  
With advanced technology nodes, resolving design rule check (DRC) violations has become a cumbersome task, which makes it desirable to make predictions at earlier stages of the design flow. In this paper, we show that the Random Forest (RF) model is quite effective for the DRC hotspot prediction at the global routing stage, and in fact significantly outperforms recent prior works, with only a fraction of the runtime to develop the model. We also propose, for the first time, to adopt a recent explanatory metric—the SHAP value—to make accurate and consistent explanations for individual DRC hotspot predictions from RF. Experiments show that RF is 21%-60% better in predictive performance on average, compared with promising machine learning models used in similar works (e.g. SVM and neural networks) while exhibiting good explainability, which makes it ideal for DRC hotspot prediction. | Download Paper (PDF; Only available from the DATE venue WiFi) |

| 18:32 | IP4-11 | **ON PRE-ASSIGNMENT ROUTE PROTOTYPING FOR IRREGULAR BUMPS ON BGA PACKAGES** |  
**Speaker:** Hung-Ming Chen, National Chiao Tung University, TW  
**Authors:** Jyun-Ru Jiang1, Yun-Chih Kuo2, Simon Chen1 and Hung-Ming Chen1  
1National Chiao Tung University, TW; 2National Taiwan University, TW; 3MediaTek,inc, TW  
**Abstract**  
In modern package design, the bumps often place irregularly due to the macros varied in sizes and positions. This will make pre-assignment routing more difficult, even with massive design efforts. This work presents a 2-stage routing method which can be applied to an arbitrary bump placement on 2-layer BGA packages. Our approach combines escape routing with via assignment: the escape routing is used to handle the irregular bumps and the via assignment is applied for improving the wire congestion and total wirelength of global routing. Experimental results based on industrial cases show that our methodology can solve the routing efficiently, and we have achieved 82% improvement on wire congestion with 5% wirelength increase compared with conventional regular treatments. | Download Paper (PDF; Only available from the DATE venue WiFi) |

| 18:30 | 8.7.4 | **SYSTEM STUDY OF SILICON PHOTONICNICS MODULATOR IN SHORT REACH GRIDLESS COHERENT NETWORKS** |  
**Speaker:** Sadok Aouini, Ciena Corporation, CA  
**Authors:** Sadok Aouini1, Ahmad Abdo1, Xueyang Li2, Md Samiul Alam2, Mahdi Parviz1, Claude D’Amours3 and David V. Plant2  
1Ciena Corporation, CA; 2McGill University, CA; 3University of Ottawa, CA  
**Abstract**  
A study the impact of modulation loss of Silicon-Photonics Mach-Zehnder modulators in the context of single-carrier coherent receivers, i.e. 400G-ZR. The modulation loss is primarily due to limited bandwidth and large peak to average ratio of the modulator output. We present the implications of performing only post-compensation of the loss at the receiver and its advantages in gridless-networks. A manageable Q factor penalty of around 0.5 dB is found for dual-polarization system with a 0.75 dB peak to average ratio (PAPR) reduction. | Download Paper (PDF; Only available from the DATE venue WiFi) |
9.2 Autonomous Systems Design Initiative: Architectures and Frameworks for Autonomous Systems

Date: Thursday 12 March 2020
Time: 08:30 - 10:00
Location / Room: Chamrousse

Chair: Selma Sadi, TU Dortmund, DE
Co-Chair: Rolf Ernst, TU Braunschweig, DE

Presentation Title: Fully Integrated Photonic Circuits on Silicon by Means of III-V/Silicon Bonding
Authors: Florian Denis-le Coarer, SCINTIL Photonics, US

Abstract: This presentation introduces a new platform integrating heterogeneous III-V/silicon gain devices at the backside of silicon-on-insulator wafers. The fabrication relies on commercial silicon photonic processes. This platform enables fully photonic integrated circuits comprising lasers, modulators, passives and photodetectors, that can be tested at the wafer level.

Presentation Title: III-V/Silicon Hybrid Lasers Integration on CMOS-Compatible 200mm and 300mm Platforms
Speaker: Karim Hassan, CEA-Leti, FR
Authors: Karim Hassan1, Siegla Bertrand1, Laetitia Aledini1, Cecilia Dupre1, Elodie Ghegin2, Philippe Rodriguez1, Fabrice Nemouchi1, Pierre Brianceau1, Antoine Schembri1, David Carrara2, Pierrick Cavallie3, Florent Franchin3, Marie-Christine Roure2, Loic Sanchez2, Christophe Jany1 and Segolene Olivier1

Abstract: We present a CMOS-compatible hybrid III-V/Silicon technology developed in CEA-Leti. Large-scale integration of silicon photonics is already available worldwide in 200mm or 300mm through different foundries, but the development of CMOS-compatible process for the III-V integration remains of major interest for next gen transceivers in the Datacom and High Performance Computing domains. The technological developments involve first the hybridization on top of a mature silicon photonic front-end wafer through direct molecular bonding, then the patterning of the III-V epitaxy layer, and then the decoration of III-V contacts through planar multilevel BEOL to be optimized. The different technological blocks will be described, and the results will be discussed on the basis of test vehicles based on either distributed feedback (DFB), distributed Bragg reflector (DBR), or Fabry-Perot (FP) laser cavities. While first demonstrations have been obtained through wafer bonding, we show that the fabrication process was subsequently validated on III-V dies bonding with a fabrication yield of Fabry-Perot lasers of 97% in 200mm. The overall technological features are expected improve the efficiency, density, and cost of silicon photonics PICs.

Presentation Title: DeepRacing: A Framework for Agile Autonomy
Speaker: Trent Weiss, University of Virginia, US
Authors: Trent Weiss and Madhur Behl, University of Virginia, US

Abstract: We consider the challenging problem of vision-based high speed autonomous racing in realistic dynamic environments. We present DeepRacing, a novel end-to-end framework, and a virtual testbed for training and evaluating algorithms for autonomous racing. The virtual testbed is implemented using the Formula One (F1) Codemasters game, which is used by many real world F1 drivers for training. We also present AdmiralNet - a Convolution Neural Network (CNN) integrated with Long Short-Term Memory (LSTM) cells that can be tuned for the autonomous racing task in the highly realistic F1 game. We evaluate AdmiralNet's performance on unseen race tracks, and also evaluate the degree of transference between the simulation and the real world by implementing end-to-end racing on a physical 1/10 scale autonomous racecar.

Download Paper (PDF; Only available from the DATE venue WiFi)

Presentation Title: Fail-Operational Automotive Software Design Using Agent-Based Graceful Degradation
Speaker: Philipp Weiss, TU Munich, DE
Authors: Philipp Weiss1, Andreas Weichslgartner2, Felix Reimann2 and Sebastian Steinhorst1
1TU Munich, DE; 2Audi Electronics Venture GmbH, DE

Download Paper (PDF; Only available from the DATE venue WiFi)

Presentation Title: A Distributed Safety Mechanism Using Middleware and Hypervisors for Autonomous Vehicles
Speaker: Pieter van der Perk, NXP Semiconductors, NL
Authors: Tjerk Bijlsma1, Andrii Buriachevskyi1, Alessandro Frigerio3, Yuting Fu4, Kees Goossens4, Ali Osman Örs2, Pieter J. van der Perk2, Andrei Terechko2 and Bart Vermeulen2
1TNO, NL; 2NXP Semiconductors, NL; 3Eindhoven University of Technology, NL; 4Eindhoven university of technology, NL

Download Paper (PDF; Only available from the DATE venue WiFi)

End of session
In-Memory Computing (IMC) represents new computing paradigm where computation happens at data location. Within the landscape of IMC approaches, non-von Neumann architectures seek to minimize data movement associated with computing. Artificial intelligence applications are one of the most promising use case of IMC since they are both compute- and memory-intensive. Running such applications on edge devices offers significant save of energy consumption and high-speed acceleration. This special session proposes to take the attendees along a journey through IMC solutions for Edge AI. This session will cover four different viewpoints of IMC for Edge AI with four talks: (i) Enabling flexible electronics very-Edge AI with IMC, (ii) design automation methodology for computational SRAM for energy efficient SIMD operations, (iii) circuit/architecture/application multiscale design and optimization methodologies for IMC architectures, and (iv) device circuit and architecture optimizations to enable PCM-based deep learning accelerators. The speakers come from three different continents (Asia, Europe, America) and four different countries (Singapore, France, USA, Switzerland). Two speakers are affiliated to academic institutes; one to industry; and one to an institute of technological research center. We strongly believe that the topic and especially selected talks are extremely hot topics in the community and will attract various people from different countries and affiliations, from both academia and industry. Furthermore, thanks to its cross layer nature, we believe that this session is tailored to touch a wide range of experts from device and circuit community up to system and application design community. We also believe that highlighting and discussing such design methodologies is a key point for high quality and high impact research.

Following up previous occurrences and success of IMC-oriented sessions and panels in DAC2019 as well as in ISLPED2019, we believe that this topic is extremely hot in the community and will trigger fruitful interactions and, we hope, collaboration among the community. We thereby expect more than 60 attendees for this session. This session will be the object of two scientific papers that will be integrated with DATE proceedings in case of acceptance.

### Time | Label | Presentation Title | Authors
--- | --- | --- | ---
08:30 | 9.3.1 | **FLEDEGE: FLEXIBLE EDGE PLATFORMS ENABLED BY IN-MEMORY COMPUTING**  
Speaker: Kamalika Datta, Nanyang Technological University, SG  
Authors: Kamalika Datta 1, Umesh Chand 2, Arko Dutt 1, Devendra Singh 2, Aaron Thean 2 and Mohamed M. Sabry 1  
1Nanyang Technological University, SG; 2National University of Singapore, SG | Download Paper (PDF; Only available from the DATE venue WiFi)

08:50 | 9.3.2 | **COMPUTATIONAL SRAM DESIGN AUTOMATION USING PUSHED-RULE BITCELLS FOR ENERGY-EFFICIENT VECTOR PROCESSING**  
Speaker: Maha Kooli, CEA-Leti, FR  
Authors: Jean-Philippe Noll, Valentin Eglolf 1, Maha Kooli 1, Roman Gauchi 1, Jean-Michel Portal 2, Henri-Pierre Charles 1, Pascal Vivet 1 and Bastien Giraud 1  
1CEA-Leti, FR; 2Aix-Marseille University, FR | Download Paper (PDF; Only available from the DATE venue WiFi)

09:10 | 9.3.3 | **DEMONSTRATING IN-CACHE COMPUTING THANKS TO CROSS-LAYER DESIGN OPTIMIZATIONS**  
Authors: Marco Rios, William Simon, Alexandre Levisse, Marina Zapater and David Atienza, EPFL, CH |  
09:35 | 9.3.4 | **DEVICE, CIRCUIT AND SOFTWARE INNOVATIONS TO MAKE DEEP LEARNING WITH ANALOG MEMORY A REALITY**  
Authors: Prithviray Narayanan, Stefano Ambrogio, Hsinyu Tsai, Katie Spoon and Geoffrey W. Burr, IBM Research, US |  
10:00 | | End of session

### 9.4 Efficient DNN design with Approximate Computing.

**Date:** Thursday 12 March 2020  
**Time:** 08:30 - 10:00  
**Location / Room:** Stendhal

**Chair:**  
Daniel Menard, INSA Rennes, FR

**Co-Chair:**  
Seokhyeong Kang, Pohang University of Science and Technology, KR

Deep Neural Networks (DNN) are widely used in numerous domains. Cross-layer DNN approximation requires efficient simulation framework. The GPU-accelerated simulation framework, ProxSim, supports DNN inference and retraining for approximate hardware. A significant amount of energy is consumed during the training process due to excessive memory accesses. The precision-controlled memory systems, dedicated for GPUs, allow flexible management of approximation. New generation of networks, like Capsule Networks, provide better learning capabilities but at the expense of high complexity. ReD-CN methodology analyzes resilience through an error injection and approximates them.

### Time | Label | Presentation Title | Authors
--- | --- | --- | ---
08:30 | 9.4.1 | **PROXSIM: SIMULATION FRAMEWORK FOR CROSS-LAYER APPROXIMATE DNN OPTIMIZATION**  
Speaker: Cecilia Eugenia De la Parra Aparicio, Robert Bosch GmbH, DE  
Authors: Cecilia De la Parra 1, Andre Guntoro 1 and Akash Kumar 2  
1Robert Bosch GmbH, DE; 2TU Dresden, DE |  
**Abstract**  
Through cross-layer approximation of Deep Neural Networks (DNN) significant improvements in hardware resources utilization for DNN applications can be achieved. This comes at the cost of accuracy degradation, which can be compensated through different optimization methods. However, DNN optimization is highly time-consuming in existing simulation frameworks for cross-layer DNN approximation, as they are usually implemented for CPU usage only. Specially for large-scale image processing tasks, the need of a more efficient simulation framework is evident. In this paper we present ProxSim, a specialized, GPU-accelerated simulation framework for approximate hardware, based on Tensorflow, which supports approximate DNN inference and retraining. Additionally, we propose a novel hardware-aware regularization technique for approximate DNN optimization. By using ProxSim, we report up to 11x savings in execution time, compared to a multi-thread CPU-based framework, and an accuracy recovery of up to 30% for three case studies of image classification with MNIST, CIFAR-10 and ImageNet. | Download Paper (PDF; Only available from the DATE venue WiFi)
ON THE AUTOMATIC EXPLORATION OF WEIGHT SHARING FOR DEEP NEURAL NETWORK COMPRESSION

Authors:
Ettienne Dupuis\textsuperscript{1}, David Novo\textsuperscript{2}, Ian O'Connor\textsuperscript{1} and Alberto Bosio\textsuperscript{1}

\textsuperscript{1}Lyon Institute of Nanotechnology, FR; \textsuperscript{2}Université de Montpellier, FR

Abstract
Deep neural networks demonstrate impressive inference results, particularly in computer vision and speech recognition. However, the computational workload and storage associated render their use prohibitive in resource-limited embedded systems. The approximate computing paradigm has been widely explored in both industrial and academic circles. It improves performance and energy-efficiency by relaxing the need for fully accurate operations. Consequently, there is a large number of implementation options with very different approximation strategies (such as pruning, quantization, low-rank factorization, knowledge distillation, ...). To the best of our knowledge, no automated approach exists for exploring, selecting and generating the best approximate versions of a given convolutional neural network (CNN) and the design objectives. The objective of this work in progress is to show that the design space exploration phase can enable significant network compression without noticeable accuracy loss. We demonstrate this via an example based on weight sharing and show that we can obtain 4x compression rate without re-training and the resulting network does not suffer from accuracy loss, in an int-16 version of LeNet-5 (5-layer, 1,720-kbit CNN) using our method.

Download Paper (PDF; Only available from the DATE venue WiFi)
The development of future memories is driven by new devices, studied to overcome the limitations of traditional memories. Among these devices STT magnetic RAMs play a fundamental role, due to their excellent performance coupled with long endurance and non-volatility. What are the issues that these memories face? How can we solve them and make them ready for a successful commercial development? And if, by changing perspective, emerging devices are used to improve existing memories like SRAM? These are some of the questions that this section aim to answer.

**Co-Chair:**
Marco Vacca, Politecnico di Torino, IT

**Time** | **Label** | **Presentation Title** | **Authors**
---|---|---|---
08:30 | 9.5.1 | IMPACT OF MAGNETIC COUPLING AND DENSITY ON STT-MRAM PERFORMANCE | Lihou Wu, TU Delft, NL
**Authors:**
Lihou Wu\(^1\), Siddharth Rao\(^2\), Mottagiallah Taouil\(^1\), Erik Jan Marinissen\(^2\), Gouri Sankar Kar\(^2\) and Said Hamdioui \(^1\)
\(^1\)TU Delft, NL; \(^2\)IMEC, BE
**Abstract**
As a unique mechanism for MRAMs, magnetic coupling needs to be accounted for when designing memory arrays. This paper models both intra- and inter-cell magnetic coupling analytically for STT-MRAMs and investigates their impact on the write performance and retention of MTJ devices, which are the data-storing elements of STT-MRAMs. We present magnetic measurement data of MTJ devices with diameters ranging from 35 nm to 175 nm, which we use to calibrate our intra-cell magnetic coupling model. Subsequently, we extrapolate this model to study inter-cell magnetic coupling in memory arrays. We propose the inter-cell magnetic coupling factor \(P_i\) to indicate coupling strength. Our simulation results show that \(P_i=2\%\) maximizes the array density under the constraint that the magnetic coupling has negligible impact on the device’s performance. Higher array densities show significant variations in average switching time, especially at low switching voltages, caused by inter-cell magnetic coupling, and dependent on the data pattern in the cell’s neighborhood. We also observe a marginal degradation of the data retention time under the influence of inter-cell magnetic coupling.

Download Paper (PDF; Only available from the DATE venue WiFi)

09:00 | 9.5.2 | HIGH-DENSITY, LOW-POWER VOLTAGE-CONTROL SPIN ORBIT TORQUE MEMORY WITH SYNCHRONOUS TWO-STEP WRITE AND SYMMETRIC READ TECHNIQUES | Wang Kang, Beihang University, CN
**Authors:**
Haotian Wang\(^1\), Wang Kang\(^1\), Liuyang Zhang\(^1\), He Zhang\(^1\), Brajesh Kumar Kaushik\(^2\) and Weisheng Zhao \(^1\)
\(^1\)Beihang University, CN; \(^2\)IIT Roorkee, IN
**Abstract**
Voltage-control spin orbit torque (VC-SOT) magnetic tunnel junction (MTJ) has the potential to achieve high-speed and low-power spintronic memory, owing to the adaptive voltage modulated energy barrier of the MTJ. However, the three-terminal device structure needs two access transistors (one for write operation and the other one for read operation) and thus occupies larger bit-cell area compared to two terminal MTJs. A feasible method to reduce area overhead is to stack multiple VC-SOT MTJs on a common antiferromagnetic strip to share the write access transistors. In this structure, high density can be achieved. However, write and read operations face problems and the design space is not sure given a strip length. In this paper, we propose a synchronous two-step multi-bit write and symmetric read method by exploiting the selective VC-SOT driven MTJ switching mechanism. Then hybrid circuits are designed and evaluated based a physics-based VC-SOT MTJ model and a 40nm CMOS design-kit to show the feasibility and performance of our method. Our work enables high-density, low-power, high-speed voltage-control SOT memory.

Download Paper (PDF; Only available from the DATE venue WiFi)

09:30 | 9.5.3 | DESIGN OF ALMOST-NONVOLATILE EMBEDDED DRAM USING NANOELECTROMECHANICAL RELAY DEVICES | Hongtao Zhong, Tsinghua University, CN
**Authors:**
Hongtao Zhong, Mingyang Gu, Juejian Wu, Huazhong Yang and Xueqing Li, Tsinghua University, CN
**Abstract**
This paper proposes low-power design of embedded dynamic random-access memory (eDRAM) using emerging nanoelectromechanical (NEM) relay devices. The motivation of this work is to reduce the standby refresh power consumption through the improvement of retention time of eDRAM cells. In this paper, it is revealed that the tunable beyond-CMOS characteristics of emerging NEM relay devices, especially the ultra-high OFF-state drain-source resistance, open up new opportunities with device-circuit co-design. In addition, the pull-in and pull-out threshold voltages are tailored to fit the operating mechanisms of eDRAM, so as to support low-voltage operations along with long retention time. Excitingly, when low-gate-leakage thick-gate transistors are used together, the proposed NEM-relay-based eDRAM exhibits so significant retention time improvement that it behaves almost "nonvolatile". Even if using thin-gate transistors in a 130nm CMOS, the evaluation of the proposed eDRAM shows up to 63x and 127x retention time improvement at 1.0V and 1.4V supply, respectively. Detailed performance benchmarking analysis, along with the practical CMOS-compatible NEM relay model, the eDRAM design and optimization considerations, is included in this paper.

Download Paper (PDF; Only available from the DATE venue WiFi)

10:00 | IP4-14 | ROBUST AND HIGH-PERFORMANCE 12-T INTERLOCKED SRAM FOR IN-MEMORY COMPUTING | Joycee Mekie, IIT Gandhinagar, IN
**Authors:**
Neelam Surana, Mili Lavanja, Abhishek Barma and Joycee Mekie, IIT Gandhinagar, IN
**Abstract**
In this paper, we analyze the existing SRAM based In-Memory Computing (IMC) proposals and show through exhaustive simulations that they fail under process variations. 6-T SRAM, 8-T SRAM, and 10-T SRAM based IMC architectures suffer from compute-disturb (stored data flips during IMC), compute-failure (provides false computation results), and half-select failures, respectively. To circumvent these issues, we propose a novel 12-T Dual Port Dual Interlocked-storage Cell (DPDICE) SRAM. DPDICE SRAM based IMC architecture (DPDICE-IMC) can perform essential boolean functions successfully in a single cycle and can perform basic arithmetic operations such as add and multiply. The most striking feature is that DPDICE-IMC architecture can perform IMC on two datasets simultaneously, thus doubling the throughput. Cumulatively, the proposed DPDICE-IMC is 26.7%, 8\times, and 28% better than 6-T SRAM, 8-T SRAM, and 10-T SRAM based IMC architectures, respectively.

Download Paper (PDF; Only available from the DATE venue WiFi)
resilient AI systems where the AI system is re-designed to tolerate critical faults or is used for error detection purposes.

Rishad Shafik, Newcastle University, GB

9.6 Intelligent Dependable Systems

Date: Thursday 12 March 2020
Time: 08:30 - 10:00
Location / Room: Lesdiguières

Chair: Rishad Shafik, Newcastle University, GB

This session spans from dependability approaches for multicore systems realized as SoCs for intelligent reliability management and on-line software-based self-test, to error resilient AI systems where the AI system is re-designed to tolerate critical faults or is used for error detection purposes.

**Time** | **Label** | **Presentation Title** | **Authors**
--- | --- | --- | ---
08:30 | 9.6.1 | THERMAL-CYCLING-AWARE DYNAMIC RELIABILITY MANAGEMENT IN MANY-CORE SYSTEM-ON-CHIP | Mohammad-Hashem Haghbayan, University of Turku, FI

Authors:
Mohammad-Hashem Haghbayan 1, Antonio Miele 2, Zhuo Zou 3, Hannu Tenhunen 1 and Juha Plosila 1
1University of Turku, FI; 2Politecnico di Milano, IT; 3Nanjing University of Science and Technology, CN

**Abstract**
Dynamic Reliability Management (DRM) is a common approach to mitigate aging and wear-out effects in multi-/many-core systems. State-of-the-art DRM approaches apply fine-grained control on resource management to increase/balance the chip reliability while considering other system constraints, e.g., performance, and power budget. Such approaches, acting on various knobs such as workload mapping and scheduling, Dynamic Voltage/Frequency Scaling (DVFS) and Per-Core Power Gating (PCPG), demonstrated to work properly with the various aging mechanisms, such as electromigration, and Negative-Bias Temperature Instability (NBTI). However, we claim that they do not suffice for thermal cycling. Thus, we here propose a novel thermal-cycling-aware DRM approach for shared-memory many-core systems running multi-threaded applications. The approach applies a fine-grained control capable at reducing both temperature levels and variations. The experimental evaluations demonstrated that the proposed approach is able to achieve 39% longer lifetime than past approaches.

Download Paper (PDF; Only available from the DATE venue WiFi)

09:00 | 9.6.2 | DETERMINISTIC CACHE-BASED EXECUTION OF ON-LINE SELF-TEST ROUTINES IN MULTI-CORE AUTOMOTIVE SYSTEM-ON-CHIPS | Andrea Floridia, Politecnico di Torino, IT

Authors:
Andrea Floridia 1, Tzammi Melendez Carmona 2, Davide Piumatti 1, AnnaChiara Russo 1, Ernesto Sanchez 1, Sergio De Luca 2, Rosario Martorana 2 and Mose Alessandro Pernici 1
1Politecnico di Torino, IT; 2STMicroelectronics, IT

**Abstract**
Traditionally, the usage of caches and deterministic execution of on-line self-test procedures have been considered two mutually exclusive concepts. At the same time, software executed in a multi-core context suffers of a limited timing predictability due to the higher system bus contention. When dealing with self-test procedures, this higher contention might lead to a fluctuating fault coverage or even the failure of some test programs. This paper presents a cache-based strategy for achieving both deterministic behaviour and stable fault coverage from the execution of self-test procedures in multi-core systems. The proposed strategy is applied to two representative modules negatively affected by a multi-core execution: synchronous imprecise interrupts logic and pipeline hazard detection unit. The experiments illustrate that it is possible to achieve a stable execution while also improving the state-of-the-art techniques for multi-core industrial System-on-Chip intended for automotive ASIL D applications.

Download Paper (PDF; Only available from the DATE venue WiFi)

09:30 | 9.6.3 | FT-CLIPACT: RESILIENCE ANALYSIS OF DEEP NEURAL NETWORKS AND IMPROVING THEIR FAULT TOLERANCE USING CLIPPED ACTIVATION | Le-Ha Hoang 1, Muhammad Abdullah Hanif 2 and Muhammad Shafique 2
1TU Wien (TU Wien), AT; 2TU Wien, AT

**Abstract**
Deep Neural Networks (DNNs) are widely being adopted for safety-critical applications, e.g., healthcare and autonomous driving. Inherently, they are considered to be highly error-tolerant. However, recent studies have shown that hardware faults that impact the parameters of a DNN (e.g., weights) can have drastic impacts on its classification accuracy. In this paper, we perform a comprehensive error resilience analysis of DNNs subjected to hardware faults (e.g., permanent faults) in the weight memory. The outcome of this analysis is leveraged to propose a novel error mitigation technique which simplifies the high-intensity fault activation values to alleviate their impact. We achieve this by replacing the unbounded activation functions with their clipped versions. We also present a method to systematically define the clipping values of the activation functions that result in increased resilience of the networks against faults. We evaluate our technique on the AlexNet and VGG-16 DNN trained for the CIFAR-10 dataset. The experimental results show that our mitigation technique significantly improves the network’s resilience to faults. For example, the proposed technique offers on average 68.92% improvement in the classification accuracy of resilience-optimized VGG-16 model at 1 x10−5 fault rate, when compared to the base network without any fault mitigation.

Download Paper (PDF; Only available from the DATE venue WiFi)
9.7 Diverse Applications of Emerging Technologies

Date: Thursday 12 March 2020
Time: 08:30 - 10:00
Location / Room: Berlioz

Chair: Pavlidis Vasilis, The University of Manchester, GB
Co-Chair: Bing Li, TU Munich, DE

This session examines a diverse set of applications for emerging technologies. Papers consider the use of Q-learning to perform more efficient backups in non-volatile processors, the use of emerging technologies to mitigate hardware side-channels, time-sequence-based classification that rise from ultrasonic patterns due to hand movements for gesture recognition, and processing-in-memory-based solutions to accelerate DNA alignment searches.

### 9.7.1 Q-LEARNING BASED BACKUP FOR ENERGY HARVESTING POWERED EMBEDDED SYSTEMS

**Speaker:** Wei Fan, Shandong University, CN
**Authors:** Wei Fan, Yujie Zhang, Weinong Song, Mengying Zhao, Zhaoyan Shen and Zhiping Jia, Shandong University, CN

**Abstract**
Non-volatile processors (NVPs) are used in energy harvesting powered embedded systems to preserve data across interruptions. In NVP systems, volatile data are backed up to non-volatile memory upon power failures and resumed after power comes back. Traditionally, backup is triggered immediately when energy warning occurs. However, it is also possible to more aggressively utilize the residual energy for program execution to improve forward progress. In this work, we propose a Q-learning based backup strategy to achieve maximal forward progress in energy harvesting powered intermittent embedded systems. The experimental results show an average of 307.4% and 43.4% improved forward progress compared with traditional instant backup and the most related work, respectively.

[Download Paper (PDF; Only available from the DATE venue WiFi)](#)

### 9.7.2 A NOVEL TIGFET-BASED DFF DESIGN FOR IMPROVED RESILIENCE TO POWER SIDE-CHANNEL ATTACKS

**Speaker:** Michael Niemier, University of Notre Dame, US
**Authors:** Mohammad Mehdi Sharifi, Ramin Rajaei, Patsy Cadareanu, Pierre-Emmanuel Gaillardon, Yier Jin, Michael Niemier and X. Sharon Hu


**Abstract**
Side-channel attacks (SCAs) represent a significant security threat, and aim to reveal otherwise secret data by analyzing a relevant circuit's behavior, e.g., its power consumption. While all circuit components are potential power side channels, D-flip-flops (DFFs) are often the primary source of information leakage to an SCA. This paper proposes a DFF design based on the three-independent-gate field-effect transistor (TIGFET) that reduces side-channel vulnerabilities of sequential circuits. Notably, we find that the I-V characteristics of the TIGFET itself leads to inherent side-channel resilience, which in turn enables simpler and more efficient cryptographic hardware. Our proposed design is based on a prior TIGFET-based true single-phase clock (TSPC) DFF design, which offers high performance and reduced area. More specifically, our modified TSPC (mTSPC) design exploits the symmetric I-V characteristics of TIGFETs, which results in pull-up and pull-down currents that are nearly identical. When combined with additional circuit modifications (made possible by the unique characteristics of the TIGFET), the mTSPC circuit draws almost the same amount of supply currents under all possible input transitions (less than 1% variation for different transitions), which can in turn mask information leakage. Using a 10nm TIGFET technology model, simulation results show that the proposed TIGFET-based DFF circuit leads to decreased power consumption (up to 96.9% when compared to the prior secured designs), has a low delay (15.2 ps), and employs only 12 TIGFET devices. Furthermore, an 8-bit S-box whose output is sampled by a group of eight mTSPC DFFs was simulated. A correlation power analysis attack on the simulated S-box with 256 power traces shows that the key is not revealed, which confirms the SCA resiliency of the proposed DFF design.

[Download Paper (PDF; Only available from the DATE venue WiFi)](#)
On the trend of ultrasound-based gesture recognition, this study introduces the concept of time-sequence classification of ultrasonic patterns induced by hand movements on a microphone array. We refer to time-sequence ultrasound echoes as continuous frequency patterns being received in real-time at different steering angles. The ultrasound source is a single tone continuously being emitted from the center of the microphone array. In the interim, the array beamforms and locates an ultrasonic activity (induced echoes) after which a processing pipeline is initiated to extract band-limited frequency features. These beamformed features are organized in a 2D matrix of size 11’x30 updated every 10ms on which a Temporal Convolutional Network (TCN) outputs continuous classification. Prior to that, the same TCN is trained to classify Doppler shift variability rate. Using this approach, we show that a user can easily achieve 49 gestures at different steering angles by means of sequence detection. To make it simple to users, we define two Doppler shift variability rates; very slow and very fast which the TCN detects 95-99% of the time. Not only a gesture can be performed at different directions but also the length of each performed gesture can be measured. This leverages the diversity of in-air ultrasound gestures allowing more control capabilities. The process is designed under low-resource settings; that is, given the fact that this real-time process is always-on, the power and memory resources should be optimized. The proposed solution needs 6.2-10.2 MMACs and a memory footprint of 6KB allowing such gesture recognition system to be hosted by energy-constrained edge devices such as smart-speakers.

9.8 Special Session: Panel: Variation-aware analyses of Mega-MOSFET Memories, Challenges and Solutions

Date: Thursday 12 March 2020
Time: 08:30 - 10:00
Location / Room: Exhibition Theatre

Moderators: Firas MOHAMED, Silvaco, FR
Jean-Baptiste DULUC, Silvaco, FR

Designing large memories under manufacturing variability requires statistical approaches that rely on SPICE simulations at different Process, Voltage, Temperature operating points to verify that yield requirements will be met. Variation-aware simulations of full memories that consist of millions of transistors is a challenging task for both SPICE simulators and statistical methodology to achieve accurate results. The ideal solution for variation-aware verifications of full memories would be to run Monte Carlo simulations through SPICE simulators to assess that all the addressable elements enable successful write and read operations. However, this classical approach suffers from practical issues and prevent it to be used. Indeed, for large memory arrays (e.g. MB and more) the number of SPICE simulations to perform would be intractable to achieve a descent statistical precision. Moreover, the SPICE simulation of a single sample of the full-memory netlist that involve millions or billions of MOSFETs and parasitic elements might be very long or impossible because of the netlist size. Unfortunately, Fast-SPICE simulations are not a palatable solution for final verification because the loss of accuracy compared to pure SPICE simulations is difficult to evaluate for such netlists. So far, most of the variation-aware methodologies to analyze and validate Mega-MOSFETs memories rely on the assumption that the sub-blocks of the system (e.g. control unit, IOs, row decoders, column circuitries, memory cells) might be assessed independently. Doing so memory designers apply dedicated statistical approaches for each individual sub-block to reduce the overall simulation time to achieve variation-aware closure. When considering that each element of the memory is independent of its neighborhood, the simulation of the memory is drastically reduced to few MOSFETs on the critical paths for read or write memory operation, the other sub-blocks being idealized and estimations being derived under Gaussian assumption. Using such an approach, memory designers avoid the usual statistical simulations of the full memory that is, most of the time, unpractical in terms of duration and load. Although the aforementioned approach has been widely used by memory designers, these methods reach their limits when designing memory for low-power and advanced-node technologies where non-idealities arise. The consequence of less reliable results is that the memory designers compensate by increasing security margins at the expense of performances to achieve satisfactory yield. In this context sub-blocks can no longer be considered individually and Gaussianity no longer prevails, other practical simulation flows are required to verify full memories with satisfying performances. New statistical approaches and simulation flows must handle memory slices or critical paths with all relevant sub-blocks in order to consider element interactions to be more realistic. Additionally, these approaches must handle the hierarchy of the memory to respect variation ranges of each sub-block, from low sigma for control units and IOs to high sigma for highly replicated blocks. Using a virtual reconstruction of the full memory the yield can be asserted without relying on the assumptions of individual sub-
block analyzes. With accurate estimation over the full memory, no more security margins are required, and better performances will be reached.”

Panelists:
- Yves Laplanche, ARM, FR
- Lorenzo Ciampolini, CEA, FR
- Pierre Faubet, SILVACO FRANCE, FR

10:00 End of session

IP4 Interactive Presentations

**Date:** Thursday 12 March 2020
**Time:** 10:00 - 10:30
**Location / Room:** Poster Area

Interactive Presentations run simultaneously during a 30-minute slot. Additionally, each IP paper is briefly introduced in a one-minute presentation in a corresponding regular session.

### Label | Presentation Title | Authors
--- | --- | ---
IP4-1 | HIT: A HIDDEN INSTRUCTION TROJAN MODEL FOR PROCESSORS | Jiaqi Zhang, Tongji University, CN; Ying Zhang, Huawei Li and Jianhui Jiang
| Speaker: | Jiaqi Zhang, Tongji University, CN |
| Authors: | Jiaqi Zhang1, Ying Zhang1, Huawei Li1 and Jianhui Jiang3 |
| **Abstract** | This paper explores an intrusion mechanism to microprocessors using illegal instructions, namely hidden instruction Trojan (HIT). It uses a low-probability sequence consisting of normal instructions as a boot sequence, followed by an illegal instruction to trigger the Trojan. The payload is a hidden interrupt to force the program counter to a specific address. Hence the program at the address has the super privileges. Meanwhile, we use integer programming to minimize the trigger probability of HIT within a given area overhead. The experimental results demonstrate that HIT has an extremely low trigger probability and can survive from the detection of the existing test methods. |
| Download Paper (PDF; Only available from the DATE venue WiFi) |

IP4-2 | BITSTREAM MODIFICATION ATTACK ON SNOW 3G | Michael Moraitis, Royal Institute of Technology KTH, SE; Elena Dubrova, Royal Institute of Technology - KTH, SE |
| Speaker: | Michael Moraitis and Elena Dubrova, Royal Institute of Technology - KTH, SE |
| Authors: | Michael Moraitis and Elena Dubrova, Royal Institute of Technology - KTH, SE |
| **Abstract** | SNOW 3G is one of the core algorithms for confidentiality and integrity in several 3GPP wireless communication standards, including the new Next Generation (NG) 5G. It is believed to be resistant to classical cryptanalysis. In this paper, we show that SNOW 3G can be broken by a fault attack based on bitstream modification. By changing the content of some look-up tables in the bitstream, we reduce the non-linear state updating function of SNOW 3G to a linear one. As a result, it becomes possible to recover the key from a known plaintext-ciphertext pair. To our best knowledge, this is the first successful bitstream modification attack on SNOW 3G. |
| Download Paper (PDF; Only available from the DATE venue WiFi) |

IP4-3 | A MACHINE LEARNING BASED WRITE POLICY FOR SSD CACHE IN CLOUD BLOCK STORAGE | Yu Zhang, Huazhong University of Science & Technology, CN; Ping Huang, Hua Wang and Jianhui Jiang |
| Speaker: | Yu Zhang, Huazhong University of Science & Technology, CN; Ping Huang, Hua Wang and Jianhui Jiang |
| Authors: | Yu Zhang1, Ke Zhou1, Ping Huang2, Hua Wang1, Jianshu Hu3, Yangtao Wang1, Yongguang Ji3 and Bin Cheng3 |
| **Abstract** | Nowadays, SSD cache plays an important role in cloud storage systems. The associated write policy, which enforces an admission control policy regarding filling data into the cache, has a significant impact on the performance of the cache system and the amount of write traffic to SSD caches. Based on our analysis on a typical cloud storage system, approximately 47.09% writes are write-only, i.e., writes to the blocks which have not been read during a certain time window. Naively writing the write-only data to the SSD cache unnecessarily introduces a large number of harmful writes to the SSD cache without any contribution to cache performance. On the other hand, it is a challenging task to identify and filter out those write-only data in a real-time manner, especially in a cloud environment running changing and diverse workloads. In this paper, to alleviate the above cache problem, we propose an ML-WP, Machine Learning Based Write Policy, which reduces write traffic to SSDs by avoiding writing write-only data. The main challenge in this approach is to identify write-only data in a real-time manner. To realize ML-WP and achieve accurate write-only data identification, we use machine learning methods to classify data into two groups (i.e., write-only and normal data). Based on this classification, the write-only data is directly written to backend storage without being cached. Experimental results show that, compared with the industry widely deployed write-back policy, ML-WP decreases write traffic to SSD cache by 41.52%, while improving the hit ratio by 2.61% and reducing the average read latency by 37.52%. |
| Download Paper (PDF; Only available from the DATE venue WiFi) |

IP4-4 | YOU ONLY SEARCH ONCE: A FAST AUTOMATION FRAMEWORK FOR SINGLE-STAGE DNN/ACCELERATOR CO-DESIGN | Weiwei Chen, Chinese Academy of Sciences, CN; Ying Wang, Shuang Yang and Lei Zhang, Chinese Academy of Sciences, CN |
| Speaker: | Weiwei Chen, Chinese Academy of Sciences, CN; Ying Wang, Shuang Yang and Lei Zhang, Chinese Academy of Sciences, CN |
| Authors: | Weiwei Chen, Ying Wang, Shuang Yang, Cheng Liu and Lei Zhang, Chinese Academy of Sciences, CN |
| **Abstract** | DNN/Accelerator co-design has shown great potential in improving QoR and performance. Typical approaches separate the design flow into two-stage: (1) designing an application-specific DNN model with high accuracy; (2) building an accelerator considering the DNN specific characteristics. However, it may fail in promising the highest composite score which combines the goals of accuracy and other hardware-related constraints (e.g., latency, energy efficiency) when building a specific neural-network-based system. In this work, we present a single-stage automated framework, YOSO, aiming to generate the optimal solution of software-and-hardware that flexibly balances between the goal of accuracy, power, and QoS. Compared with the two-stage method on the baseline systolic array accelerator and Cifar10 dataset, we achieve 1.42x~2.29x energy or 1.79x~3.07x latency reduction at the same level of precision, for different user-specified energy and latency optimization constraints, respectively. |
| Download Paper (PDF; Only available from the DATE venue WiFi) |
**EXPLFRAME: EXPLOITING PAGE FRAME CACHE FOR FAULT ANALYSIS OF BLOCK CIPHERS**

**Speaker:**
Anirban Chakraborty, IIT Kharagpur, IN

**Authors:**
Anirban Chakraborty¹, Sarani Bhattacharya², Sayandeep Saha¹ and Debdeep Mukhopadhyay¹

**IIT Kharagpur, IN;** Scuola Superiore Sant’Anna / Università di Pisa, IT; ¹IIT Kharagpur, IN; ²KU Leuven, BE

**Abstract**

Page Frame Cache (PFC) is a purely software cache, present in modern Linux based operating systems (OS), which stores the page frames that were recently released by the processes running on a particular CPU. In this paper, we show that the page frame cache can be maliciously exploited using a vendor-provided standard cell library and commercial timing analysis tools. By reducing the worst-case output skew by over 70%, the test case example was able to achieve equivalent throughput of an 8-staged sequentially pipelined implementation with power savings of almost 3X.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

**WHEN SORTING NETWORK MEETS PARALLEL BITSTREAMS: A FAULT-TOLENT PARALLEL TERNARY NEURAL NETWORK ACCELERATOR BASED ON STOCHASTIC COMPUTING**

**Speaker:**
Yawen Zhang, Peking University, CN

**Authors:**
Yawen Zhang¹, Sheng Lin², Runsheng Wang¹, Yanzhi Wang², Yuan Wang¹, Weikang Qian³ and Ru Huang¹

¹Peking University, CN; ²Northeastern University, US; ³Shanghai Jiao Tong University, CN

**Abstract**

Stochastic computing (SC) has been widely used in neural networks (NNs) due to its simple hardware cost and high fault tolerance. Conventionally, SC-based NN accelerators adopt a hybrid stochastic-binary format, using an accumulative parallel counter to convert bitstreams into a binary number. This method, however, sacrifices the fault tolerance and causes a high hardware cost. In order to fully exploit the superior fault tolerance of SC, taking a ternary neural network (TNN) as an example, we propose a parallel SC-based NN accelerator purely using bitstream computation. We apply a bitonic sorting network for simultaneously implementing the accumulation and activation function with parallel bitstreams. The proposed design not only has high fault tolerance, but also achieves at least 2.8 energy efficiency improvement over the binary computing counterpart.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

**EFFICIENT EMBEDDED MACHINE LEARNING APPLICATIONS USING ECHO STATE NETWORKS**

**Speaker:**
Rolando Brondolin, Politecnico di Milano, IT

**Authors:**
Luca Cerina¹, Giuseppe Franco², Claudio Gallicchio³, Alessio Micheli³ and Marco D. Santambrogio⁴

¹Ipolitecnico di milano, Italy; ²Scuola Superiore Sant’Anna / Università di Pisa, IT; ³Università di Pisa, IT; ⁴Politecnico di Milano, IT

**Abstract**

The increasing role of Artificial Intelligence (AI) and Machine Learning (ML) in our lives brought a paradigm shift on how and where the computation is performed. Stringent latency requirements and congested bandwidth moved AI inference from Cloud space towards end-devices. This change required a major simplification of Deep Neural Networks (DNN), with memory-wise libraries or co-processors that perform fast inference with minimal power. Unfortunately, many applications such as natural language processing, time-series analysis and audio interpretation are built on a different type of Artificial Neural Networks (ANN), the so-called Recurrent Neural Networks (RNN), which, due to their intrinsic architecture, remains too complex and heavy to run efficiently on embedded devices. To solve this issue, the Reservoir Computing paradigm proposes sparse untrained non-linear networks, the Reservoir, that can embed temporal relations without some of the hindrances of Recurrent Neural Networks (RNN), and with a lower memory usage. Echo State Networks (ESN) and Liquid State Machines are the most notable examples. In this scenario, we propose a performance comparison of an ESN, designed and trained using Bayesian Optimization techniques, against current RNN solutions. We aim to demonstrate that ESN have comparable performance in terms of accuracy, require minimal training time, and they are more optimized in terms of memory usage and computational efficiency. Preliminary results show that ESN are competitive with RNN on a simple benchmark, and both training and inference time are faster, with a maximum speed-up of 2.35x and 6.60x, respectively.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

**WAVEPRO: CLOCK-LESS WAVE-PROPEGATED PIPELINE COMPILER FOR LOW-POWER AND HIGH-THROUGHPUT COMPUTATION**

**Speaker:**
Yeohuda Kra, Bar-Ilan University, IL

**Authors:**
Yeohuda Kra, Adam Teman and Tzachi Noy, Bar-Ilan University, IL

**Abstract**

Clock-less Wave-Propagated Pipelining is a longknown approach to achieve high-throughput without the overhead of costly sampling registers. However, due to many design challenges, which have only increased with technology scaling, this approach has never been widely accepted and has generally been limited to small and very specific demonstrations. This paper addresses this barrier by presenting WavePro, a generic and scalable algorithm, capable of skew balancing any combinatorial logic netlist for the application of wave pipelining. The algorithm was implemented in the WavePro Compiler automation utility, which interfaces with industry delays extraction and standard timing analysis tools to produce a sign-off quality result. The utility is demonstrated upon a dot-product accelerator in a 65 nm CMOS technology, using a vendor-provided standard cell library and commercial timing analysis tools. By reducing the worstcase output skew by over 70%, the test case example was able to achieve equivalent throughput of an 8-staged sequentially pipelined implementation with power savings of almost 3X.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

**DEEPNVM: A FRAMEWORK FOR MODELING AND ANALYSIS OF NON-VOLATILE MEMORY TECHNOLOGIES FOR DEEP LEARNING APPLICATIONS**

**Speaker:**
Ahmet Inci, Carnegie Mellon University, US

**Authors:**
Ahmet Inci¹, Mehmet M Ig Nec⁴, and Diana Marculescu², Carnegie Mellon University, US

¹Carnegie Mellon University, US; ²University of Illinois at Urbana-Champaign, US; ³UC Berkeley, CA; ⁴Scuola Superiore Sant'Anna, Italy

**Abstract**

Non-volatile memory (NVM) technologies such as spin-transfer torque magnetic random access memory (STT-MRAM) and spin-orbit torque magnetic random access memory (SOT-MRAM) have significant advantages compared to conventional SRAM due to their non-volatility, higher cell density, and scalability features. While previous work has investigated several architectural implications of NVM for generic applications, in this work we present DeepNVM, a framework to characterize, model, and analyze NVM-based caches in GPU architectures for deep learning (DL) applications by combining technology-specific circuit-level models and the actual memory behavior of various DL workloads. We present both iso-capacity and iso-area performance and energy analysis for systems whose last-level caches rely on conventional SRAM and emerging STT-MRAM and SOT-MRAM technologies. In the iso-capacity case, SOT-MRAM and STT-MRAM provide up to 4.2x and 5x energy-delay product (EDP) reduction and 2.4x and 3x area reduction compared to conventional SRAM, respectively. Under iso-area assumptions, SOT-MRAM and STT-MRAM provide 2x EDP reduction on average across all workloads when compared to SRAM. Our comprehensive cross-layer framework is demonstrated on STT-/SOT-MRAM technologies and can be used for the characterization, modeling, and analysis of any NVM technology for last-level caches in GPU platforms for deep learning applications.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

**WHEN SORTING NETWORK MEETS PARALLEL BITSTREAMS: A FAULT-TOLENT PARALLEL TERNARY NEURAL NETWORK ACCELERATOR BASED ON STOCHASTIC COMPUTING**

**Speaker:**
Anirban Chakraborty, IIT Kharagpur, IN

**Authors:**
Anirban Chakraborty¹, Sarani Bhattacharya², Sayandeep Saha¹ and Debdeep Mukhopadhyay¹

¹IIT Kharagpur, IN; ²Shanghai Jiao Tong University, CN

**Abstract**

Stochastic computing (SC) has been widely used in neural networks (NNs) due to its simple hardware cost and high fault tolerance. Conventionally, SC-based NN accelerators adopt a hybrid stochastic-binary format, using an accumulative parallel counter to convert bitstreams into a binary number. This method, however, sacrifices the fault tolerance and causes a high hardware cost. In order to fully exploit the superior fault tolerance of SC, taking a ternary neural network (TNN) as an example, we propose a parallel SC-based NN accelerator purely using bitstream computation. We apply a bitonic sorting network for simultaneously implementing the accumulation and activation function with parallel bitstreams. The proposed design not only has high fault tolerance, but also achieves at least 2.8 energy efficiency improvement over the binary computing counterpart.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

**EFFICIENT EMBEDDED MACHINE LEARNING APPLICATIONS USING ECHO STATE NETWORKS**

**Speaker:**
Rolando Brondolin, Politecnico di Milano, IT

**Authors:**
Luca Cerina¹, Giuseppe Franco², Claudio Gallicchio³, Alessio Micheli³ and Marco D. Santambrogio⁴

¹Ipolitecnico di milano, Italy; ²Scuola Superiore Sant’Anna / Università di Pisa, IT; ³Università di Pisa, IT; ⁴Politecnico di Milano, IT

**Abstract**

The increasing role of Artificial Intelligence (AI) and Machine Learning (ML) in our lives brought a paradigm shift on how and where the computation is performed. Stringent latency requirements and congested bandwidth moved AI inference from Cloud space towards end-devices. This change required a major simplification of Deep Neural Networks (DNN), with memory-wise libraries or co-processors that perform fast inference with minimal power. Unfortunately, many applications such as natural language processing, time-series analysis and audio interpretation are built on a different type of Artificial Neural Networks (ANN), the so-called Recurrent Neural Networks (RNN), which, due to their intrinsic architecture, remains too complex and heavy to run efficiently on embedded devices. To solve this issue, the Reservoir Computing paradigm proposes sparse untrained non-linear networks, the Reservoir, that can embed temporal relations without some of the hindrances of Recurrent Neural Networks (RNN), and with a lower memory usage. Echo State Networks (ESN) and Liquid State Machines are the most notable examples. In this scenario, we propose a performance comparison of an ESN, designed and trained using Bayesian Optimization techniques, against current RNN solutions. We aim to demonstrate that ESN have comparable performance in terms of accuracy, require minimal training time, and they are more optimized in terms of memory usage and computational efficiency. Preliminary results show that ESN are competitive with RNN on a simple benchmark, and both training and inference time are faster, with a maximum speed-up of 2.35x and 6.60x, respectively.

Download Paper (PDF; Only available from the DATE venue WiFi)
[IP4-10] XGBIR: AN XGBOOST-BASED IR DROP PREDICTOR FOR POWER DELIVERY NETWORK

Speaker: An-Yu Su, National Chiao Tung University, TW
Authors: Chi-Hsien Pao, Yu-Min Lee and An-Yu Su, National Chiao Tung University, TW
Abstract
This work utilizes the XGBoost to build a machine-learning-based IR drop predictor, XGBIR, for the power grid. To capture the behavior of power grid, we extract its several features and employ its locality property to save the extraction time. XGBIR can be effectively applied to large designs and the average error of predicted IR drops is less than 6 mV.

Download Paper (PDF; Only available from the DATE venue WiFi)

[IP4-12] TOWARDS BEST-EFFORT APPROXIMATION: APPLYING NAS TO APPROXIMATE COMPUTING

Speaker: Weiwei Chen, Chinese Academy of Sciences, CN
Authors: Weiwei Chen, Ying Wang, Shuang Yang, Cheng Liu and Lei Zhang, Chinese Academy of Sciences, CN
Abstract
The design of neural network architecture for code approximation involves a large number of hyper-parameters to explore, it is a non-trivial task to find an neural-based approximate computing solution that meets the demand of application-specified accuracy and Quality of Service (QoS). Prior works do not address the problem of ‘optimal’ network architectures design in program approximation, which depends on the user-specified constraints, the complexity of dataset and the hardware configuration. In this paper, we apply Neural Architecture Search (NAS) for searching and selecting the neural approximate computing and provide an automatic framework that tries to generate the best-effort approximation-result while satisfying the user-specified QoS/accuracy constraints. Compared with previous method, this work achieves more than 1.4x speedup and 1.7x energy reduction on average when applied to the AxBench benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

[IP4-13] ON THE AUTOMATIC EXPLORATION OF WEIGHT SHARING FOR DEEP NEURAL NETWORK COMPRESSION

Speaker: Etienne Dupuis, École Centrale de Lyon, FR
Authors: Etienne Dupuis, David Novo, Ian O'Connor and Alberto Bossio
Abstract
Deep neural networks demonstrate impressive inference results, particularly in computer vision and speech recognition. However, the computational workload and storage associated render their use prohibitive in resource-limited embedded systems. The approximate computing paradigm has been widely explored in both industry and academic circles. It improves performance and energy-efficiency by relaxing the need for fully accurate operations. Consequently, there is a large number of implementation options with very different approximation strategies (such as pruning, quantization, low-rank factorization, knowledge distillation, ...). To the best of our knowledge, no automated approach exists for exploring, selecting and generating the best approximate versions of a given convolutional neural network (CNN) and the design objectives. The objective of this work in progress is to show that the design space exploration phase can enable significant network compression without noticeable accuracy loss. We demonstrate this via an example based on weight sharing and show that we can obtain 4x compression rate without re-training and the resulting network does not suffer from accuracy loss, in an int-16 version of LeNet-5 (5-layer, 1,720-kbit CNN) using our method.

Download Paper (PDF; Only available from the DATE venue WiFi)

[IP4-14] ROBUST AND HIGH-PERFORMANCE 12-T INTERLOCKED SRAM FOR IN-MEMORY COMPUTING

Speaker: Joycee Mekie, IIT Gandhinagar, IN
Authors: Neelam Surana, Mili Lavania, Abhishek Barma and Joycee Mekie, IIT Gandhinagar, IN
Abstract
In this paper, we analyze the existing SRAM based In-Memory Computing (IMC) proposals and show through exhaustive simulations that they fail under process variations. 6-T SRAM, 8-T SRAM, and 10-T SRAM based IMC architectures suffer from compute-disturb (stored data flips during IMC), compute-failure (provides false computation results), and half-select failures, respectively. To circumvent these issues, we propose a novel 12-T Dual Port Dual Interlocked-storage Cell (DPDICE) SRAM. DPDICE SRAM based IMC architecture (DPDICE-IMC) can perform essential boolean functions successfully in a single cycle and can perform basic arithmetic operations such as add and multiply. The most striking feature is that DPDICE-IMC architecture can perform IMC on two datasets simultaneously, thus doubling the throughput. Cumulatively, the proposed DPDICE-IMC is 26.7%, 8 times, and 28% better than 6-T SRAM, 8-T SRAM, and 10-T SRAM based IMC architectures, respectively.

Download Paper (PDF; Only available from the DATE venue WiFi)

[IP4-15] HIGH DENSITY SRT-MRAM COMPILER DESIGN, VALIDATION AND CHARACTERIZATION METHODOLOGY IN 28NM FDSOI TECHNOLOGY

Speaker: Piyush Jain, ARM Embedded Technologies Pvt. Ltd., IN
Authors: Piyush Jain, Akshay Kumar, Nicolaas Van Winkelhoff, Didier Gayraud, Surya Gupta, Abdelali El Amrouzi, Giorgio Palmis, Alexandra Gourio, Laurentz Vachez, Luc Palsus, Jean-Christophe Buy and Cyrille Dray
Abstract
Spin Transfer Torque Magnetic-resistive Random-Access Memory (SRT-MRAM) is emerging as a promising substitute for flash memories due to scaling challenges for flash in process nodes beyond 28nm. SRT-MRAM’s high endurance, fast speed and low power makes it suitable for wide variety of applications. An embedded MRAM (eMRAM) compiler is highly desirable to enable SoC designers to use eMRAM instances in their designs in a flexible manner. However, the development of an eMRAM compiler has added challenges of handling multi-fold higher density and maintaining analog circuits accuracy, on top of the challenges associated with conventional SRAM memory compilers. In this paper, we present a successful design methodology for a high density 128Mb eMRAM compiler in a 28nm fully depleted SOI (FDSOI) process. This compiler enables optimized eMRAM instance generation with varying capacity ranges, word-widths, and optional features like repair and error correction. eMRAM compiler design is achieved by evolving various architecture design, validations and characterization methods. A hierarchical and modular characterization methodology is proposed to enable high accuracy characterization and industry-standard EDA view generation from the eMRAM compiler.

Download Paper (PDF; Only available from the DATE venue WiFi)
AN APPROXIMATION-BASED FAULT DETECTION SCHEME FOR IMAGE PROCESSING APPLICATIONS

Speaker:
Antonio Miele, Politecnico di Milano, IT

Authors:
Matteo Biasielli, Luca Cassano and Antonio Miele, Politecnico di Milano, IT

Abstract:
Image processing applications expose an intrinsic resilience to faults. In this application field the classical Duplication with Comparison (DWC) scheme, where output images are discarded as soon as the two replicas' outputs differ for at least one pixel, may be over-conservative. This paper introduces a novel lightweight fault detection scheme for image processing applications: i) it extends the DWC scheme by substituting one of the two exact replicas with a faster approximated one; and ii) it features a Neural Network-based checker designed to distinguish between usable and unusable images instead of faulty/unfaulty ones. The application of the hardening scheme on a case study has shown an execution time reduction from 27% to 34% w.r.t. the DWC, while guaranteeing a comparable fault detection capability.

Download Paper (PDF; Only available from the DATE venue WiFi)

TRANSPORT-FREE MODULE BINDING FOR SAMPLE PREPARATION USING MICROFLUIDIC FULLY PROGRAMMABLE VALVE ARRAYS

Speaker:
Gautam Choudhary, Adobe Research, India, IN

Authors:
Gautam Choudhary1, Sandeep Patil1, Debraj Kundu1, Sukanta Bhattacharjee1, Shigeru Yamashita1, Bing Li4, Ulf Schlichtmann4 and Sudip Roy1

IIT Roorkee, IN; 2Indian Statistical Institute, IN; 3Ritsumeikan University, JP; 4TU Munich, DE

Abstract:
Microfluidic fully programmable valve array (FPVA) biochips have emerged as general-purpose flow-based microfluidic lab-on-chips (LoCs). An FPVA supports highly re-configurable on-chip components (modules) in the two-dimensional grid-like structure controlled by some software programs, unlike application-specific flow-based LoCs. Fluids can be loaded into or washed from a cell with the help of flows from the inlet to outlet of an FPVA, whereas cell-to-cell transportation of discrete fluid segments is not precisely possible. The simplest mixing module to realize on an FPVA-based LoC is a four-way mixer consisting of a $2 \times 2$ array of cells working as a ring-like mixer having four valves. In this paper, we propose a design automation method for sample preparation that finds suitable placements of mixing operations of a mixing tree using four-way mixers without requiring any transportation of fluid(s) between modules. We also propose a heuristic that modifies the mixing tree to reduce the sample preparation time. We have performed an extensive simulation and examined several parameters to determine the performance of the proposed solution.

Download Paper (PDF; Only available from the DATE venue WiFi)

TAPASCO: THE OPEN-SOURCE TASK-PARALLEL SYSTEM COMPOSER FRAMEWORK

Authors:
Carsten Heinze, Lukas Sommer, Lukas Weber, Jaco Hofmann and Andreas Koch, TU Darmstadt, DE

Abstract:
Field-programmable gate arrays (FPGA) are an established platform for highly specialized accelerators, but in a heterogeneous setup, the accelerator still needs to be integrated into the overall system. The open-source TaPaSCo (Task-Parallel System Composer) framework was created to serve this purpose: The fast integration of FPGA-based accelerators into compute platforms or systems-on-chip (SoC) and their connection to relevant components on the FPGA board. TaPaSCo can support developers in all steps of the development process: from cores resulting from High-Level Synthesis or cores written in an HDL, a complete FPGA-design can be created. TaPaSCo will automatically connect all processing elements to the memory- and host-interface and generate a complete bitstream. The TaPaSCo Runtime API allows to interface with accelerators from software and supports operations such as transferring data to the FPGA memory, passing values and controlling the execution of the accelerators.

More information ...

TAPUSFI: PARTICLE FILTER FUSION ASIC FOR INDOOR POSITIONING

Authors:
Christian Schott, Marko Rößler, Daniel Froß, Marcel Putsche and Ulrich Heinke, TU Chemnitz, DE

Abstract:
The meaning of data acquired from IoT devices is heavily enhanced if global or local position information of their acquirement is known. Infrastructure for indoor positioning as well as the IoT device involve the need of small, energy efficient but powerful devices that provide the location awareness. We propose the PAFUSI, a hardware implementation of an UWB position estimation algorithm that fulfills these requirements. Our design fuses distance measurements to fixed points in an environment to calculate the position in 3D space and is capable of using different positioning technologies like GPS, DecaWave or Nanotron as data source simultaneously. Our design comprises of an estimator which processes the data by means of a Sequential Monte Carlo method and a microcontroller core which configures and controls the measurement unit as well as it analyses the results of the estimator. The PAFUSI is manufactured as a monolithic integrated ASIC in a H2020-MSCA-ITN RESCUE European Union project.

More information ...
More information ...
FU: LOW POWER AND ACCURACY CONFIGURABLE APPROXIMATE ARITHMETIC UNITS

Authors:
Tomoaki Ukezono and Toshinori Sato, Fukuoka University, JP

Abstract
In this demonstration, we will introduce the approximate arithmetic units such as adder, multiplier, and MAC that are being studied in our system-architecture laboratory. Our approximate arithmetic units can reduce delay and power consumption at the expense of accuracy. Our approximate arithmetic units are intended to be applied to IoT edge devices that can process images, and are suitable for battery-driven and low-cost devices. The biggest feature of our approximate arithmetic units is that the circuit is configured so that the accuracy is dynamically variable, and the trade-off relationship between accuracy and power can be selected according to the usage status of the device. In this demonstration, we show the power consumption according to various accuracy-requirements based on actual data and claim the practicality of the proposed arithmetic units.

More information ...

10.1 Special Day on “Silicon Photonics”: High-Speed Silicon Photonics Interconnects for Data Center and HPC

Date: Thursday 12 March 2020
Time: 11:00 - 12:30
Location / Room: Amphithéâtre Jean Prouve

Chair:
Ian O’Connor, Ecole Centrale de Lyon, FR

Co-Chair:
Luca Ramini, Hewlett Packard Labs, US

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00</td>
<td>10.1.1</td>
<td>THE NEED AND CHALLENGES OF CO-PACKAGING AND OPTICAL INTEGRATION IN DATA CENTERS</td>
</tr>
<tr>
<td>Author:</td>
<td>Luca Gantz, Mellanox, US</td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td>Silicon photonic (SiPh) technology was the &quot;talk of the town&quot; for almost two decades, yet only in the last couple of years, actual SiPh based transceivers were introduced for short-reach links. As the global IP traffic skyrockets, and will surpass 1 ZB per year by 2020, it seems that this is the optimal point for new disruptive technology to emerge. SiPh technology has the potential to reduce power consumption while meeting the demand for increasing rates, and potentially even reduce the cost. Yet in order to fully integrate SiPh components in mainly CMOS ICs, the entire industry must align, beginning with industrial FABs and OSATs, and ending with system manufacturers and Data Center clients. Indeed, in the last year positive developments have occurred as the Hyper-scalers are starting to show interest in driving the market into integrating optics and forgo pluggable transceivers. Yet many challenges have to be met, and some hard decisions have to be taken in order to fully integrate optics in a scalable manner. In this talk I will review these challenges and possible ways to meet them in order to enable optical integrated products in Data Centers and High-Performance Computers.</td>
<td></td>
</tr>
</tbody>
</table>

| 11:30 | 10.1.2 | POWER AND COST ESTIMATE OF SCALABLE ALL-TO-ALL TOPOLOGIES WITH SILICON PHOTONICS LINKS |
| Author: | Luca Ramini, Hewlett Packard Labs, US |
| Abstract | For many applications that require a tight latency profile, such as machine learning, a network topology that does not leverage arbitration-based switching is desired. All-to-all (A2A) interconnection networks enable any node in the network to communicate to any other node at any given time. Many abstractions can be made to enable this capability such as buffering, time-domain multiplexing, etc. However, typical A2A topologies are limited to about 32 nodes within one hop. This is primarily due to limitations in reach, power consumption and bandwidth per interconnect. In this presentation, a topology of 256 nodes and beyond is considered by leveraging the many-wavelengths-per-fiber advantage of DWDM silicon photonics technology. Power and cost estimate of scalable A2A topologies using silicon photonics links are provided in order to understand the practical limits, if any, of a single node communicating with many other nodes via one wavelength per node. |

| 12:00 | 10.1.3 | THE NEXT FRONTIER IN SILICON PHOTONIC DESIGN: EXPERIMENTALLY VALIDATED STATISTICAL MODELS |
| Author: | Geoff Duggan1, James Pond1, Xu Wang2, Ellen Schelew1, Federico Gomez1, Milad Mahpeykar1, Ray Chung1, Zequin Lu1, Parya Samadian1, Jens Niegemann1, Adam Reid1, Roberto Armenta1, Dylan McGuire1, Peng Sun2, Jared Hulme2, Mudit Jan1 and Ashkan Seyedi2 |
| Lumerical, US; 2Hewlett Packard Labs, US |
| Abstract | Silicon photonics has made tremendous progress in recent years and is now a critical technology embedded in many commercial products, particularly for data communications, while new products in sensing, AI and even quantum information technologies are in development. High quality processes from multiple foundries, supported by sophisticated electronic-photon design automation (EPDA) workflows have made these advancements possible. Although several initiatives have begun to address the issue of manufacturing variability in photonics, these approaches have not been integrated meaningfully into EPDA workflows which lag well behind electronic integrated circuit workflows. Contributing to this deficiency has been a lack of data to calibrate statistical photonic compact models used in photonics circuit and system simulation. We present our current work in developing tools to calibrate statistical photonic compact models and compare our results against experimental data. |

12:30 End of session


Date: Thursday 12 March 2020
Time: 11:00 - 12:30
Location / Room: Chamrousse

Chair:
Philipp Mundhenk, Autonomous Intelligent Driving GmbH, DE

Co-Chair:
Ahmad Adee, Bosch Corporate Research, DE

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00</td>
<td>End of session</td>
<td></td>
</tr>
</tbody>
</table>
In this special session, we bring academia and industry together to discuss latest results and future directions for research in the domain of next-generation computer arithmetic.

754-2008 compliant arithmetic. Optimizations on fixed-point and integer arithmetic are also pursued actively for low-power computing architectures. Furthermore, approximate advances in the domain of computer arithmetic like high-precision anchored numbers from ARM, posit arithmetic by John Gustafson, and bfloat16, etc. as an alternative to IEEE Arithmetic is ubiquitous in today's digital world, ranging from embedded to high-performance computing systems. With machine learning at fore in a wide range of application domains from wearables, automotive, avionics to weather prediction, sufficiently accurate yet low-cost arithmetic is the need for the day. Recently, there have been several advances in the domain of computer arithmetic like high-precision anchored numbers from ARM, posit arithmetic by John Gustafson, and bfloat16, etc. as an alternative to IEEE 754-2008 compliant arithmetic. Optimizations on fixed-point and integer arithmetic are also pursued actively for low-power computing architectures. Furthermore, approximate computing and transprecision/mixed-precision computing have been exciting areas for research forever. While academic research in the domain of computer arithmetic has a long history, industrial adoption of some of these new data-types and techniques is in its early stages and expected to increase in future. bfloat16 is an excellent example of that.

The rise of machine learning pushes the massive compute power requirements, especially on the edge devices for their real-time inferences. One established approach for reducing the power usage is by going down to integer inferences (such as 8-bit) instead of utilizing higher computation accuracy given by their floating-point counterparts. Squeezing into lower bit representations such as in binary weight networks or binary neural networks requires complex training methods and also more efforts to recover the precision loss, and it typically functions only on simple classification tasks. One promising alternative to further reduce power consumption and computation latency is by utilizing approximate compute units. This method is a promising paradigm for mitigating the computation demand of neural networks, by taking advantage of their inherent resilience. Thanks to the development in approximate computing in the last decade, we have abundant options to utilize the best available approximate units, without re-developing or re-designing them. Nonetheless, adaptation during training phase is required. At first, we need to adapt the training methods for neural networks to take into account the inaccuracy given by approximate compute, without sacrificing the training speed (considering the trainings are performed on GPU with floating-point). Second, we need to define new metric for assessing and selecting the best-fit of approximation units per use-case basis. Lastly, we need to take advantages of approximation into the neural networks, such as over-fitting mitigation per design and resiliency, so that the networks trained for and designed with approximation will and shall perform better than their exact computing counterparts. For these steps, we evaluate on small tasks first and further validate on complex tasks which are more relevant in automotive domains.
The recent FPGA architectures have introduced new levels of embedded floating point performance, with tens of TFLOPs now available across a wide range of device sizes. The last two generations of FPGAs have introduced IEEE754 single precision (FP32) arithmetic, containing up to 10 TFLOPs. The emergence of AI/Machine Learning as the highest profile FPGA application has changed the focus from signal processing and embedded calculations supported by FP32 to smaller floating point precisions, such as BFLOAT16 for training and FP16 for inference. This talk, we will describe the architecture and development of the Intel Agilex DSP Block, which contains a FP32 multiplier-adder pair that can be decomposed into two smaller precision pairs: fp16, bfloat16, and a third proprietary format which can be used for both training and inference. In the Edge, where even lower precision arithmetic is required for inference, new FPGA EDA flows can implement 100 TFLOPs+ of soft logic-based compute power. In the second half of our talk, we will describe new synthesis, clustering, and packing methodologies - collectively known as Fractal Synthesis - that allow an unprecedented near 100% logic use of the FPGA for arithmetic, while maintaining the clock rates of a small example design. The soft logic and embedded arithmetic capabilities can be used simultaneously, making the FPGA the most flexible, and amongst the highest performing AI platform.

12:06  10.3.4 A COMPARISON OF POSIT AND IEEE 754 FLOATING-POINT ARITHMETIC THAT ACCOUNTS FOR EXCEPTION HANDLING

Author: John Gustafson, National University of Singapore, SG

Abstract: The posit number format has advantages over the decades-old IEEE 754 Standard floating-point standard in many dimensions: Accuracy, dynamic range, simplicity, bitwise reproducibility, resiliency, and resistance to side-channel security attacks. In making comparisons, it is essential to distinguish between an IEEE 754 Standard implementation that handles all the exceptions in hardware, and one that either ignores the exceptions of the Standard or handles them with software or microcode that take hundreds of clock cycles to execute. Ignoring the exceptions quickly leads to egregious problems such as different values comparing as equal; handling exceptions with microcode creates massive data-dependency on timing that permits side-channel attacks like the well-known Spectre and Meltdown security weaknesses. Many microprocessors, such as current x86 architectures, use the exception-trapping approach for exceptions such as denormalized floats, which makes them unsuitable for secure use. Post arithmetic provides data-independent and fast execution times with less complexity than a data-independent IEEE 754 float environment for the same data size.

End of session

10.4 Design Methodologies for Hardware Approximation

Date: Thursday 12 March 2020
Time: 11:00 - 12:30
Location / Room: Stendhal

Chair: Lukas Sekanina, Brno University of Technology, CZ
Co-Chair: David Novo, CNRS & University of Montpellier, FR

New methods for the design and evaluation of approximate hardware are key to its success. This section shows that these approximation methods are applicable across different levels of hardware description including an RTL design of an approximate multiplier, approximate circuits modelled using binary decision diagrams and a behavioural description used in the context of high level synthesis of hardware accelerators. The papers of this section also show how to address another challenge - an efficient error evaluation - by means of new statistical and formal verification methods.

11:00  10.4.1 REALM: REDUCED-ERROR APPROXIMATE LOG-BASED INTEGER MULTIPLIER

Speaker: Hassan Saadat, University of New South Wales, AU
Authors: Hassan Saadat, Haris Javaid, Aleksandar Ignjatovic and Sri Parameswaran

Abstract: We propose a new error-configurable approximate unsigned integer multiplier named REALM. It incorporates a novel error-reduction method into the classical approximate log-based multiplier. Each power-of-two-interval of the input operands is partitioned into MxN segments, and an error-reduction factor for each segment is analytically determined. These error-reduction factors can be used across any power-of-two-interval, so we quantize only M^2 factors and store them in the form of read-only hardwired lookup tables to keep the resource overhead to a minimum. Error characterization of REALM shows that it achieves very low error bias (mostly less than or equal to 0.05%), along with lower mean error (from 0.4% to 1.6%), and lower peak error (from 2.08% to 7.4%) than the classical approximate log-based multiplier and its state-of-the-art derivatives (mean errors greater than or equal to 2.6% and peak errors greater than or equal to 7.8%). Synthesis results using TSMC 45nm standard-cell library show that REALM enables significant power-efficiency (66% to 86% reduction) and area-efficiency (50% to 76% reduction) when compared with the accurate integer multiplier. We show that REALM produces Pareto optimal design trade-offs in the design space of state-of-the-art approximate multipliers. Application-level evaluation of REALM demonstrates that it has negligible effect on the output quality.

Download Paper (PDF; Only available from the DATE venue WiFi)
### 10.4 Emerging Machine Learning Applications and Models

#### 10:30 10.4.2 A FAST BDD MINIMIZATION FRAMEWORK FOR APPROXIMATE COMPUTING

**Speaker:** Oliver Kesozoce, Friedrich-Alexander-Universität Erlangen-Nürnberg, DE  
**Authors:** Andreas Wendler and Oliver Kesozoce, Friedrich-Alexander-Universität Erlangen-Nürnberg, DE  
**Abstract**  
Approximate Computing is a design paradigm that trades off computational accuracy for gains in non-functional aspects such as reduced area, increased computation speed, or power reduction. Computing the error of the approximated design is an essential step to determine its quality. The computation time for determining the error can become very large, effectively rendering the entire logic approximation procedure infeasible. As a remedy, we present methods to accelerate the computation of error metric computations by (a) exploiting structural information and (b) computing estimates of the metrics for multi-output Boolean functions represented as BDDs. We further present a novel greedy, bucket-based BDD minimization framework employing the newly proposed error metric computations to produce Pareto-optimal solutions with respect to BDD size and multiple error metrics. The applicability of the proposed minimization framework is demonstrated by an experimental evaluation. We can report considerable speedups while, at the same time, creating high-quality approximated BDDs.  
**Download Paper (PDF; Only available from the DATE venue WiFi)**

#### 10:40 10.4.3 ON THE DESIGN OF HIGH PERFORMANCE HW ACCELERATOR THROUGH HIGH-LEVEL SYNTHESIS SCHEDULING APPROXIMATIONS

**Speaker:** Benjamin Carrion Schaefer, University of Texas at Dallas, US  
**Authors:** Siyuan Xu and Benjamin Carrion Schaefer, University of Texas at Dallas, US  
**Abstract**  
High-level synthesis (HLS) takes as input a behavioral description (e.g., C/C++) and generates efficient hardware through three main steps: allocation, scheduling, and binding. The scheduling step, times the operations in the behavioral description by scheduling different portions of the code at unique clock steps (control steps). The code portions assigned to each clock step mainly depend on the target synthesis frequency and target technology. This work makes use of this to generate smaller and faster circuits by approximating the program portions scheduled in each clock step and by exploiting the slack between different scheduling steps to further increase the performance/reduce the latency of the resultant circuit. In particular, each individual scheduling step is approximated given a maximum error boundary and a library of different approximation techniques. In order to further optimize the resultant circuit, different scheduling steps are merged based on the timing slack of different control step without violating the given timing constraint (target frequency). Experimental results from different domain-specific applications show that our method works well and is able to increase the throughput on average by 82% while at the same time reducing the area by 21% for a given maximum allowable error.  
**Download Paper (PDF; Only available from the DATE venue WiFi)**

#### 10:50 10.4.4 FAST KRIGING-BASED ERROR EVALUATION FOR APPROXIMATE COMPUTING SYSTEMS

**Speaker:** Daniel Menard, INSA Rennes, FR  
**Authors:** Justine Bonnot1, Karol Desnos1 and Daniel Menard2  
1Université de Rennes / Inria / IRISA, FR; 2INSA Rennes, FR  
**Abstract**  
Approximate computing techniques trade-off the performance of an application for its accuracy. The challenge when implementing approximate computing in an application is to efficiently evaluate the quality at the output of the application to optimize the noise budgeting of the different approximation sources. It is commonly achieved with an optimization algorithm to minimize the implementation cost of the application subject to a quality constraint. During the optimization process, numerous approximation configurations are tested, and the quality at the output of the application is measured for each configuration with simulations. The optimization process is a time-consuming task. We propose a new method for inferring the accuracy or quality metric at the output of an application using kriging, a geostatistical method.  
**Download Paper (PDF; Only available from the DATE venue WiFi)**

### 10.5 Emerging Machine Learning Applications and Models

**Date:** Thursday 12 March 2020
This session presents new application domains and new models for neural networks, discussing two novel video applications: multi-view and surveillance, and discussing a Bayesian model approach for neural networks.

**11:00** 10.5.1 COMMUNICATION-EFFICIENT VIEW-POOLING FOR DISTRIBUTED INFERENCE WITH MULTI-VIEW NEURAL NETWORKS

**Speaker:** Manik Singhal, School of Electrical and Computer Engineering, Purdue University, US

**Authors:** Manik Singhal, Anand Raghunathan and Vijay Raghunathan, Purdue University, US

**Abstract**

Multi-view object detection or the problem of detecting an object using multiple viewpoints, is an important problem in computer vision with varied applications such as distributed smart cameras and collaborative drone swarms. Multi-view object detection algorithms based on deep neural networks (DNNs) achieve high accuracy by (em view pooling), or aggregating features corresponding to the different views. However, when these algorithms are realized on networks of edge devices, the communication cost incurred by view pooling often dominates the overall latency and energy consumption. In this paper, we propose techniques for communication-efficient view pooling that can be used to improve the efficiency of distributed multi-view object detection and apply them to state-of-the-art multi-view DNNs. First, we propose (em significance-aware selective view pooling), which identifies and communicates only those features from each view that are likely to impact the pooled result (and hence, the final output of the DNN). Second, we propose (em multi-resolution feature view pooling), which divides views into dominant and non-dominant views, and down-scales the features from non-dominant views using an additional network layer before communicating them for pooling. The dominant and non-dominant views are pooled separately and the results are jointly used to derive the final classification. We implement and evaluate the proposed pooling schemes using a model test-bed of twelve Raspberry Pi 3b+ devices and show that they achieve 9×-36× reduction in data communicated and 1.8× reduction in inference latency, with no degradation in accuracy.

Download Paper (PDF; Only available from the DATE venue WiFi)

**11:30** 10.5.2 AN ANOMALY COMPREHENSION NEURAL NETWORK FOR SURVEILLANCE VIDEOS ON TERMINAL DEVICES

**Speaker:** Yuan Cheng, Shanghai Jiao Tong University, CN

**Authors:** Yuan Cheng, Guangtai Huang, Peining Zhen, Bin Liu, Hai-Bao Chen, Nga Wong and Hao Yu

**Abstract**

Anomaly comprehension in surveillance videos is more challenging than detection. This work introduces the design of a lightweight and fast anomaly comprehension neural network. For comprehension, a spatio-temporal LSTM model is developed based on the structured, tensorized time-series features extracted from surveillance videos. Deep compression of network size is achieved by tensorization and quantization for the implementation on terminal devices. Experiments on large-scale video anomaly dataset UCF-Crime demonstrate that the proposed network can achieve an impressive inference speed of 266 FPS on a GTX-1080Ti GPU, which is 4.29 faster than ConvLSTM-based method; a 3.34% AUC improvement with 5.55% accuracy niche versus the 3D-CNN based approach; and at least 15k× parameter reduction and 228× storage compression over the RNN-based approaches. Moreover, the proposed framework has been realized on an ARM-core based IOT board with only 2.4W power consumption.

Download Paper (PDF; Only available from the DATE venue WiFi)

**12:00** 10.5.3 BYNQNNet: BAYESIAN NEURAL NETWORK WITH QUADRATIC ACTIVATIONS FOR SAMPLING-FREE UNCERTAINTY ESTIMATION ON FPGA

**Speaker:** Hiromitsu Awano, Osaka University, JP

**Authors:** Hiromitsu Awano and Masanori Hashimoto, Osaka University, JP

**Abstract**

An efficient inference algorithm for Bayesian neural network (BNN) named BYNQNNet, Bayesian neural network with quadratic activations, and its FPGA implementation are proposed. As neural networks find applications in mission critical systems, uncertainty estimations in network inference become increasingly important. BNN is a theoretically grounded solution to deal with uncertainty in neural network by treating network parameters as random variables. However, an inference in BNN involves Monte Carlo (MC) sampling, i.e., a stochastic forwarding is repeated N times with randomly sampled network parameters, which results in N times slower inference compared to non-Bayesian approach. Although recent papers proposed sampling-free algorithms for BNN inference, they still require evaluation of complex functions such as a cumulative distribution function (CDF) of Gaussian distribution for propagating uncertainties through nonlinear activation functions such as ReLU and Heaviside, which requires considerable amount of resources for hardware implementation. Contrary to conventional BNN, BYNQNNet employs quadratic nonlinear activation functions and hence the uncertainty propagation can be achieved using only polynomial operations. Our numerical experiment reveals that BYNQNNet has comparative accuracy with MC-based BNN which requires N=10 forwardings. We also demonstrate that BYNQNNet implemented on Xilinx PYQ-Z1 FPGA board achieves the throughput of 131x10^3 images per second and the energy efficiency of 44.7×10^3 images per joule, which corresponds to 4.07x and 8.99x improvements from the state-of-the-art MC-based BNN accelerator.

Download Paper (PDF; Only available from the DATE venue WiFi)

12:30 End of session

**10.6 Secure Processor Architecture**

**Date:** Thursday 12 March 2020

**Time:** 11:00 - 12:30

**Location / Room:** Lesdiguières

**Chair:** Emanuele Regnath, TU Munich, DE

**Co-Chair:** Erkay Savas, Sabanci University, TR

This session proposes an overview of new mechanisms to protect processor architectures, boot sequences, caches, and energy management. The solutions strive to address and mitigate a wide range of attack methodologies, with a special focus on new emerging attacks.

Download Paper (PDF; Only available from the DATE venue WiFi)
11:00 10.6.1 CAPTURING AND OBSCURING PING-PONG PATTERNS TO MITIGATE CONTINUOUS ATTACKS
Speaker:
Kai Wang, Harbin Institute of Technology, CN
Authors:
Kai Wang1, Fengkai Yuan2, Rui Hou2, Zhenzhou Ji1 and Dan Meng2
1Harbin Institute of Technology, CN; 2Chinese Academy of Sciences, CN
Abstract
In this paper, we observed Continuous Attacks are one kind of common side channel attack scenarios, where an adversary frequently probes the same target cache lines in a short time. Continuous Attacks cause target cache lines to go through multiple load-evict processes, exhibiting Ping-Pong Patterns. Identifying and obscuring Ping-Pong Patterns effectively interferes with the attacker’s probe and mitigates Continuous Attacks. Based on the observations, this paper proposes Ping-Pong Regulator to identify multiple Ping-Pong Patterns and block them with different strategies (Preload or Lock). The Preload proactively loads target lines into the cache, causing the attacker to mistakenly infer that the victim has accessed these lines; the Lock fixes the attacked lines’ directory entries on the last level cache directory until they are evicted out of caches, making an attacker’s observation of the locked lines is always the L2 cache miss. The experimental evaluation demonstrates that the Ping-Pong Regulator efficiently identifies and secures attacked lines, induces negligible performance impacts and storage overhead, and does not require any software support.
Download Paper (PDF; Only available from the DATE venue WiFi)

11:30 10.6.2 MITIGATING CACHE-BASED SIDE-CHANNEL ATTACKS THROUGH RANDOMIZATION: A COMPREHENSIVE SYSTEM AND ARCHITECTURE LEVEL ANALYSIS
Speaker:
Houman Homayoun, University of California, Davis, US
Authors:
Han Wang1, Hossein Sayadi1, Avesta Sasan1, Setareh Rafatirad1, Houman Homayoun1, Liang Zhao1 and Tinoosh Mohsenin2
1George Mason University, US; 2University of Maryland, Baltimore County, US
Abstract
Cache hierarchy was designed to allow CPU cores to process instructions faster by bridging the significant latency gap between the main memory and processor. In addition, various cache replacement algorithms are proposed to predict future data and instructions to boost the performance of the computer systems. However, recently proposed cache-based SideChannel Attacks (SCAs) have shown to effectively exploiting such a hierarchical cache design. The cache-based SCAs are exploiting the hardware vulnerabilities to steal secret information from users by observing cache access patterns of cryptographic applications and thus are emerging as a serious threat to the security of the computer systems. Prior works on mitigating the cache-based SCAs have mainly focused on cache partitioning techniques and/or randomization of mapping between hardware main memory and shared cache memory. However, such solutions though effective, require modification in the processor hardware which increases the complexity of architecture design and are not applicable to current architectures. In response, this paper proposes a lightweight system and architecture level randomization technique to effectively mitigate the impact of side-channel attacks on last-level caches with no hardware redesign overhead for current as well as legacy architectures. To aim, by carefully adapting the processor frequency and prefsetters operation and adding proper level of noise to the attackers’ cache observations we attempt to protect the critical information from being leaked. The experimental results indicate that the concurrent randomization of frequency and.prefsetters can significantly prevent cache-based side-channel attacks with no need for a new cache design. In addition, the proposed randomization and adaptation methodology outperform state-of-the-art solutions in terms of the performance and execution time by reducing the performance overhead from 32.66% to nearly 20%.
Download Paper (PDF; Only available from the DATE venue WiFi)

12:00 10.6.3 EXTENDING THE RISC-V INSTRUCTION SET FOR HARDWARE ACCELERATION OF THE POST-QUANTUM SCHEME LAC
Speaker:
Tim Fritzmann, TU Munich, DE
Authors:
Tim Fritzmann1, Georg Sigl1 and Johanna Sepúlveda3
1TU Munich, DE; 2TU Munich/Fraunhofer ASIEC, DE; 3Airbus Defence and Space, DE
Abstract
The increasing effort in the development of quantum computers represents a high risk for communication systems due to their capability of breaking currently used public-key cryptography. LAC is a lattice-based public-key encryption scheme resistant to traditional and quantum attacks. It is characterized by small key sizes and low arithmetic complexity. Recent publications have shown practical post-quantum solutions through co-design techniques. However, for LAC only software implementations were explored. In this work, we propose an efficient, flexible and time-protected HW/SW co-design architecture for LAC. We present two contributions. First, we develop and integrate hardware accelerators for three LAC performance bottlenecks: the generation of polynomials, polynomial multiplication and error correction. The accelerators were designed to support all post-quantum security levels from 128 to 256-bits. Second, we develop tailored instruction set extensions for LAC on RISC-V and integrate the HW accelerators directly into a RISC-V core. The results show that our architecture for LAC with constant-time error correction improves the performance by a factor of 7.66 for LAC-128, 14.42 for LAC-192, and 13.36 for LAC-256, when compared to the unprotected reference implementation running on RISC-V. The increased performance comes at a cost of an increased cache-based SCAs exploiting the hardware vulnerabilities to steal secret information from users by observing cache access patterns of cryptographic applications and thus are emerging as a serious threat to the security of the computer systems. Prior works on mitigating the cache-based SCAs have mainly focused on cache partitioning techniques and/or randomization of mapping between hardware main memory and shared cache memory. However, such solutions though effective, require modification in the processor hardware which increases the complexity of architecture design and are not applicable to current architectures. In response, this paper proposes a lightweight system and architecture level randomization technique to effectively mitigate the impact of side-channel attacks on last-level caches with no hardware redesign overhead for current as well as legacy architectures. To aim, by carefully adapting the processor frequency and prefsetters operation and adding proper level of noise to the attackers’ cache observations we attempt to protect the critical information from being leaked. The experimental results indicate that the concurrent randomization of frequency and prefsetters can significantly prevent cache-based side-channel attacks with no need for a new cache design. In addition, the proposed randomization and adaptation methodology outperform state-of-the-art solutions in terms of the performance and execution time by reducing the performance overhead from 32.66% to nearly 20%.
Download Paper (PDF; Only available from the DATE venue WiFi)

12:30 10.7 Accelerators for Neuromorphic Computing
Date: Thursday 12 March 2020
Time: 11:00 - 12:30
Location / Room: Berloz
Chair: Alexandre Levisse, EPFL, CH
Co-Chair:
**Slide 1**

In this session, special hardware accelerators based on different technologies for neuromorphic computing will be presented. These accelerators (i) improve the computing efficiency by using pulse widths to deliver information across memristor crossbars, (ii) enhance the robustness of neuromorphic computing with unary coding and priority mapping, and (iii) explore the modulation of light in transferring information so to push the performance of computing systems to new limits.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00</td>
<td>10.7.1</td>
<td><strong>A PULSE WIDTH NEURON WITH CONTINUOUS ACTIVATION FOR PROCESSING-IN-MEMORY ENGINES</strong></td>
<td>Shuhang Zhang, TU Munich, DE</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Speaker:</strong> Shuhang Zhang, Bing Li, Hai (Helen) Li and Ulf Schlichtmann 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Authors:</strong> Shuhang Zhang, Bing Li, Hai (Helen) Li and Ulf Schlichtmann 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1TU Munich, DE; 2Duke University, US / TU Munich, US</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Processing-in-memory engines have been applied successfully to accelerate deep neural networks. For improving computing efficiency, spiking-based platforms are widely utilized. However, spiking-based designs quantize inter-layer signals naturally, leading to performance loss. In addition, the spike mismatch effect makes digital processing an essential part, impeding direct signal transferring between layers and thus resulting in longer latency. In this paper, we propose a novel neuron design based on pulse width modulation, avoiding quantization step and bypassing spike mismatch via its continuous activation. The computation latency and circuit complexity can be reduced significantly due to the absence of quantization and digital processing steps, while keeping a competitive performance. Experimental results demonstrate that the proposed neuron design can achieve &gt;100× speedup, and the area and power consumption can be reduced up to 75% and 25% compared with spiking-based designs.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
<td></td>
</tr>
<tr>
<td>11:30</td>
<td>10.7.2</td>
<td><strong>GO UNARY: A NOVEL SYNAPSE CODING AND MAPPING SCHEME FOR RELIABLE RERAM-BASED NEUROMORPHIC COMPUTING</strong></td>
<td>Li Jiang, Shanghai Jiao Tong University, CN</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Speaker:</strong> Chang Ma, Yanan Sun, Weikang Qian, Ziqi Meng, Rui Yang and Li Jiang, Shanghai Jiao Tong University, CN</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Authors:</strong> Chang Ma, Yanan Sun, Weikang Qian, Ziqi Meng, Rui Yang and Li Jiang, Shanghai Jiao Tong University, CN</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Neuroval network (NN) computing contains a large number of multiply-and-accumulate (MAC) operations, which is the speed bottleneck in traditional von Neumann architecture. Resistive random access memory (ReRAM)-based crossbars are well suited for matrix-vector multiplication. Existing ReRAM-based NNs are mainly based on the binary coding for synaptic weights. However, the imperfect fabrication process combined with stochastic filament-based switching leads to resistance variations, which can significantly affect the weights in binary synapses and degrade the accuracy of NNs. Further, as multi-level cells (MLCs) are being developed for reducing hardware overhead, the NN accuracy deteriorates more due to the resistance variations in the binary coding. In this paper, a novel unary coding of synaptic weights is presented to overcome the resistance variations of MLCs and achieve reliable ReRAM-based neuromorphic computing. The priority mapping is also proposed in compliance with the unary coding to guarantee high accuracy by mapping those bits with lower resistance states to ReRAMs with smaller resistance variations. Our experimental results show that the proposed method provides less than 0.45% and 5.48% accuracy loss on LeNet (on MNIST dataset) and VGG16 (on CIFAR-10 dataset), respectively, while maintaining acceptable hardware cost.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
<td></td>
</tr>
<tr>
<td>12:00</td>
<td>10.7.3</td>
<td><strong>LIGHTBULB: A PHOTONIC-NONVOLATILE-MEMORY-BASED ACCELERATOR FOR BINARIZED CONVOLUTIONAL NEURAL NETWORKS</strong></td>
<td>Farzaneh Zokaei1, Qian Lou1, Nathan Youngblood2, Weichen Liu3, Yiyan Xie4 and Lei Jiang1</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Authors:</strong> Farzaneh Zokaei1, Qian Lou1, Nathan Youngblood2, Weichen Liu3, Yiyan Xie4 and Lei Jiang1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1Indiana University Bloomington, US; 2University of Pittsburgh, US; 3Wanyang Technological University, SG; 4Southwest University, CN</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Although Convolutional Neural Networks (CNNs) have demonstrated the state-of-the-art inference accuracy in various intelligent applications, each CNN inference involves millions of expensive floating point multiply-accumulate (MAC) operations. To efficiently process CNN inferences, prior work proposes an electro-optical accelerator to process power-of-2 quantized CNNs by electro-optical ripple-carry adders and optical binary shifters. The electro-optical accelerator also uses SRAM registers to store intermediate data. However, electro-optical ripple-carry adders and SRAMs seriously limit the operating frequency and inference throughput of the electro-optical accelerator, due to the long critical path of the adder and the long access latency of SRAMs. In this paper, we propose a photonics random access memory (NVM)-based accelerator, LightBulb, to process binarized CNNs by high frequency photonic XNOR gates and popcount units. LightBulb also adopts photonic racetrack memory to serve as input/output registers to achieve high operating frequency. Compared to prior electro-optical accelerators, on average, LightBulb improves the CNN inference throughput by $17\times$ and $31\times$ and $5\times$ respectively, while maintaining acceptable hardware cost.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td>10.7.4</td>
<td><strong>ROQ: A NOISE-AWARE QUANTIZATION SCHEME TOWARDS ROBUST OPTICAL NEURAL NETWORKS WITH LOW-BIT CONTROLS</strong></td>
<td>Jiaqi Gu, University of Texas at Austin, US</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Speaker:</strong> Jiaqi Gu, Zheng Zhao1, Chenghao Feng2, Hanqing Zhu3, Ray T. Chen1 and David Z. Pan1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1University of Texas at Austin, US; 2Shanghai Jiao Tong University, CN</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Optical neural networks (ONNs) demonstrate orders-of-magnitude higher speed in deep learning acceleration than their electronic counterparts. However, limited control precision and device variations induce accuracy degradation in practical ONN implementations. To tackle this issue, we propose a quantization scheme that adapts a full-precision ONN to low-resolution voltage controls. Moreover, we propose a protective regularization technique that dynamically penalizes quantized weights based on their estimated noise-robustness, leading to an improvement in noise robustness. Experimental results show that the proposed scheme effectively adapts ONNs to limited-precision controls and does not sacrifice performance. The results show a four-layer ONN demonstrates higher inference accuracy with lower variances than baseline methods under various control precisions and device noises.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td>10.7.5</td>
<td><strong>STATISTICAL TRAINING FOR NEUROMORPHIC COMPUTING USING MEMRISTOR-BASED CROSSBARS CONSIDERING PROCESS VARIATIONS AND NOISE</strong></td>
<td>Ying Zhu, TU Munich, DE</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Speaker:</strong> Ying Zhu1, Grace Li Zhang1, Tianchen Wang2, Bing Li1, Yiyuan Xie4 and Lei Jiang1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1TU Munich, DE; 2University of Notre Dame, US; 3National Tsing Hua University, TW</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Abstract</strong> Memristor-based crossbars are an attractive platform to accelerate neuromorphic computing. However, process variations during manufacturing and noise in memristors cause significant accuracy loss if not addressed. In this paper, we propose to model process variations and noise as correlated random variables and incorporate them into the cost function during training. Consequently, the weights after this statistical training become more robust and together with global variation compensation provide a stable inference accuracy. Simulation results demonstrate that the mean value and the standard deviation of the inference accuracy can be improved significantly, by even up to 54% and 31%, respectively, in a two-layer fully connected neural network.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
<td></td>
</tr>
</tbody>
</table>

**End of session**
**UB10 Session 10**

**Date:** Thursday 12 March 2020  
**Time:** 12:00 - 14:30  
**Location / Room:** Booth 11, Exhibition Area

---

**Label** | **Presentation Title** | **Authors**
---|---|---

**UB10.1**  
**TAPASCO: THE OPEN-SOURCE TASK-PARALLEL SYSTEM COMPOSER FRAMEWORK**  
**Authors:** Carsten Heinz, Lukas Sommer, Lukas Weber, Jacy Hofmann and Andreas Koch, TU Darmstadt, DE

**Abstract**
Field-programmable gate arrays (FPGA) are an established platform for highly specialized accelerators, but in a heterogeneous setup, the accelerator still needs to be integrated into the overall system. The open-source TaPaSCo (Task-Parallel System Composer) framework was created to serve this purpose: The fast integration of FPGA-based accelerators into compute platforms or systems-on-chip (SoC) and their connection to relevant components on the FPGA board. TaPaSCo can support developers in all steps of the development process: from cores resulting from High-Level Synthesis or cores written in an HDL, a complete FPGA-design can be created. TaPaSCo will automatically connect all processing elements to the memory- and host-interface and generate a complete bitstream. The TaPaSCo Runtime API allows to interface with accelerators from software and supports operations such as transferring data to the FPGA memory, passing values and controlling the execution of the accelerators.

**More information ...**

**UB10.2**  
**RESCUED: A RESCUE DEMONSTRATOR FOR INTERDEPENDENT ASPECTS OF RELIABILITY, SECURITY AND QUALITY TOWARDS A COMPLETE EDA FLOW**  
**Authors:** Nevin George1, Guilherme Cardoso Medeiros 2, Junchao Chen3, Josie Esteban Rodríguez Condia4, Thomas Lange5, Aleksa Damjanovic6, Raphaël Segabianzi7, Ferreira1, Aneesh Balakrishnan5, Xinhuai Li8, Shayanesh Masoomian9, Dmytro Petryk10, Troya Cagil Koylu11, Felipe Augusto da Silva12, Ahmet Cagri Bagbab13, Cemil Cem Gürsöy14, Said Hamdioui15, Motiqaiahla Taouil16, Milos Krstic17, Peter Langendoerfer18, Zoya Dyka19, Marcelo Brandalero1, Michael Hübner1, Jörg Noite1, Theodor Vierhaus1, Matteo Sonza Reorda20, Giovanni Squillero21, Luca Sterpone22, Jaan Raik23, Dan Alexandrescu24, Maximilien Glioreux25, Georgios Selimis26, Geert-Jan Schijn27, Anton Klotz28, Christian Sauer29 and Maximilian Jenihin30

1Brandenburg University of Technology Cottbus-Senftenberg, DE; 2TU Delft, NL; 3Leibniz-Institut für innovative Mikroelektronik, DE; 4Politecnico di Torino, IT; 5IROC Technologies, FR; 6Tallinn University of Technology, EE; 7Intrinsic ID, NL; 8Cadence Design Systems GmbH, DE

**Abstract**
Demonstrator highlights the various interdependent aspects of Reliability, Security and Quality in nanoelectronics system design within an EDA toolset and a processor architecture setup. The compelling need of attention towards these three aspects of nanoelectronic systems have been ever more pronounced over extreme miniaturization of technologies. Further, such systems have exploded in numbers with IoT devices, heavy and analogous interaction with the external physical world, complex safety-critical applications, and Artificial intelligence applications. RESCUE targets such aspects in the form, Reliability (functional safety, ageing, soft errors), Security (tamper-resistance, PUF technology, intelligent security) and Quality (novel fault models, functional test, FMEA/MEA). Verification/debug spanning the entire hardware software system stack. The demonstrator is brought together by a group of PhD students under the banner of H2020-MSCA-ITN RESCUE European Union project.

**More information ...**

**UB10.3**  
**RETINE: A PROGRAMMABLE 3D STACKED VISION CHIP ENABLING LOW LATENCY IMAGE ANALYSIS**  
**Authors:** Stéphane Chevbœ`be1, Maria Lepecq2 and Laurent Millet2

1CEA LIST, FR; 2C2E-Letli, FR

**Abstract**
We have developed and fabricated a 3D stacked imager called RETINE composed with 2 layers based on the replication of a programmable 3D tile in a matrix manner providing a highly parallel programmable architecture. This tile is composed by a 16×16 851 binned pixels array with associated readout and 16 column ADC on the first layer coupled to an efficient SIMD processor of 16 PE on the second layer. The prototype of RETINE achieves high video rates, from 550 fps in binned mode to 340 fps in full resolution mode. It operates at 80 MHz with 720 mW power consumption leading to 85 GOPS/W power efficiency. To highlight the capabilities of the RETINE chip we have developed a demonstration platform with an electronic board embedding a RETINE chip that films rotating disks. Three scenario are available: high speed image capture, slow motion and composed image capture with parallel processing during acquisition.

**More information ...**

**UB10.4**  
**DL PUF ENAU: DEEP LEARNING BASED PHYSICALLY UNCLONABLE FUNCTION ENROLLMENT AND AUTHENTICATION**  
**Authors:** Amir Alipour1, David Hely2, Vincent Berouillé2 and Giorgio Di Natale3

1Grenoble INP, FR; 2Grenoble INP, FR; 3CNRS / Grenoble INP / TIMA, FR

**Abstract**
Physically Unclonable Functions (PUFs) have been addressed nowadays as a potential solution to improve the security in authentication and encryption process in Cyber Physical Systems. The research on PUF is actively growing due to its potential of being secure, easily implementable and expandable, using considerably less energy. To use PUF in common, the low level device Hardware Variation is captured per unit for device enrollment into a format called Challenge-Response Pair (CRP), and recaptured after device is deployed, and compared with the original for authentication. These enrollment + comparison functions can vary and be more complex safety-critical applications, and Artificial intelligence applications. RESCUE targets such aspects in the form, Reliability (functional safety, ageing, soft errors), Security (tamper-resistance, PUF technology, intelligent security) and Quality (novel fault models, functional test, FMEA/MEA). Verification/debug spanning the entire hardware software system stack. The demonstrator is brought together by a group of PhD students under the banner of H2020-MSCA-ITN RESCUE European Union project.

**More information ...**

**UB10.5**  
**LAGARTO: FIRST SILICON RISC-V ACADEMIC PROCESSOR DEVELOPED IN SPAIN**  
**Authors:** Guillen Cabo Pitarch1, Cristobal Ramirez Lazo1, Julian Pavon Rivera1, Vatistas Kostalabros1, Carlos Rojas Morales1, Miquel Moreto1, Jaume Abella1, Francisco J. Cazorla1, Adrian Cristal1, Roger Figueras2, Alberto Gonzalez2, Carles Hernandez2, Cesar Hernandez2, Neili Leyva2, Joan Marino1, Ricardo Martinez3, Iñaki Mendizabal4, Francesc Moll,1 Marco Antonio Ramirez2, Carlos Rojas4, Antonio Rubio4, Abraham Ruiz4, Nehir Sonmez1, Lluis Teres3, Osman Unsal5, Mateo Valero1, Ivan Vargas1 and Luis Villa2

1BSC / UPC, ES; 2CIC-IPN, MX; 3IMB-CNIC (CSIC), ES; 4UPC, ES; 5BSC, ES

**Abstract**
Open hardware is a possibility that has emerged in recent years and has the potential to be as disruptive as Linux was once, an open source software paradigm. If Linux managed to lessens the dependency of users in large companies providing software and software applications, it is envisioned that hardware based on ISAs open source can do the same in their own field. In the Lagarto tapeout four research institutions were involved: Centro de Investigación en Computación del Mexican IPN, Centro Nacional de Microelectrónica of the CSIC, Universitat Politècnica de Catalunya (UPC) and Barcelona Supercomputing Center (BSC). As a result, many bachelor, master and PhD students had the chance to achieve real-world experience with ASIC design and achieve a functional SoC. In the booth, you will find a live demo of the first ASIC and prototypes running on FPGA of the next versions of the SoC core.

**More information ...**
UB10.6 SRSN: SECURE RECONFIGURABLE TEST NETWORK
Authors: Vincent Reynaud†, Emanuele Valea†, Paolo Maistri†, Regis Leveugle†, Marie-Lise Flottes†, Sophie Dupuis‡, Bruno Rouzeyre‡ and Giorgio Di Natale†
TIMA Laboratory, FR; ‡LIRMM, FR
Abstract
The critical importance of testability for electronic devices led to the development of IEEE test standards. These methods, if not protected, offer a security backdoor to attackers. This demonstrator illustrates a state-of-the-art solution that prevents unauthorized usage of the test infrastructure based on the IEEE 1687 standard and implemented on an FPGA target.
More information ...

UB10.7 DEEPSENSE-FPGA: FPGA ACCELERATION OF A MULTIMODAL NEURAL NETWORK
Authors: Mehdi Trabelsi Ajili and Yuko Hara-Azumi, Tokyo Institute of Technology, JP
Abstract
Currently, Internet of Things and Deep Learning (DL) are merging into one domain and creating outstanding technologies for various classification tasks. Such technologies require complex DL networks that are mainly targeting powerful platforms with rich computing resources like servers. Therefore, for resource-constrained embedded systems, new challenges of size, performance and power consumption have to be considered, particularly when edge devices handle multimedial data, i.e., different types of real-time sensing data (voice, video, text, etc.). Our ongoing project is focused on DeepSense, a multimedial DL framework combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to process time-series data, such as accelerometer and gyroscope to detect human activity. We aim at accelerating DeepSense by FPGA (Xilinx Zynq) in a hardware-software co-design manner. Our demo will show the latest achievements through latency and power consumption evaluations.
More information ...

UB10.8 GENERATING ASYNCHRONOUS CIRCUITS FROM CATAPULT
Authors: Yoan Decoudu†, Jean Simatic†, Katell Morin-Allory† and Laurent Fesquet†
†University Grenoble Alpes, FR; ‡HawAI.Tech, FR; †Université Grenoble Alpes, FR
Abstract
In order to spread asynchronous circuit design to a large community of designers, High-Level Synthesis (HLS) is probably a good choice because it requires limited design technical skills. HLS usually provides an RTL description, which includes a data-path and a control-path. The desynchronization process is only applied to the control-path, which is a Finite State Machine (FSM). This method is sufficient to make asynchronous the circuit. Indeed, data are processed step by step in the pipeline stages, thanks to the desynchronized FSM. Thus, the data-path computation time is no longer related to the clock period but rather to the average time for processing data into the pipeline. This tends to improve speed when the pipeline stages are not well-balanced. Moreover, our approach helps to quickly designing data-driven circuits while maintaining a reasonable cost, a similar area and a short time-to-market.
More information ...

UB10.9 PA-HLS: HIGH-LEVEL ANNOTATION OF ROUTING CONGESTION FOR XILINX VIVADO HLS DESIGNS
Authors: Osama Bin Tariq1, Junnan Shan1, Luciano Lavagno1, Georgios Floros2, Mihai Teodor Lazarescu1, Christos Sotiriou2 and Mario Roberto Casu1
1Politecnico di Torino, IT; 2University of Thessaly, GR
Abstract
We will demo a novel high-level backannotation flow that reports routing congestion issues at the C++ source level by analyzing reports from FPGA physical design (Xilinx Vivado) and internal debugging files of the Vivado HLS tool. The flow annotates the C++ source code, identifying likely causes of congestion, e.g., on-chip memories or the DSP units. These shared resources often cause routing problems on FPGAs because they cannot be duplicated by physical design. We demonstrate on realistic large designs how the information provided by our flow can be used to both identify congestion issues at the C++ source level and solve them using HLS directives. The main demo steps are: 1-Extraction of the source-level debugging information from the Vivado HLS database 2-Generation of a list of net names involved in congestion areas and of their relative significance from the Vivado post global-routing database 3-Visualization of the C++ code lines that contribute most to congestion
More information ...

UB10.10 ATECES: AUTOMATED TESTING THE ENERGY CONSUMPTION OF EMBEDDED SYSTEMS
Author: Eduard Enoiu, Mälardalen University, SE
Abstract
The demonstrator will focus on automatically generating test suites by selecting test cases using random test generation and mutation testing is a solution for improving the efficiency and effectiveness of testing. Specifically, we generate and select test cases based on the concept of energy-aware mutants, small syntactic modifications in the system architecture, intended to mimic real energy faults. Test cases that can distinguish a certain behavior from its mutations can increase the fault detection ability. This kind of results should motivate both academia and industry to investigate the use of automatic test generation for energy consumption.
More information ...

14:30 End of session

11.0 LUNCHTIME KEYNOTE SESSION
Date: Thursday 12 March 2020
Time: 12:30 - 13:50
Location / Room: Amphithéâtre Jean Prouvé
Chair: Gabriela Nicolaeu, Polytechnique Montréal, CA
Co-Chair: Luca Ramini, Hewlett Packard Labs, US

Time Label Presentation Title Authors
**11.1 Special Day on “Silicon Photonics”: Advanced Applications**

**Date:** Thursday 12 March 2020  
**Time:** 14:00 - 15:30  
**Location / Room:** Amphithéâtre Jean Prouve

**Chair:** Olivier Sentieys, University of Rennes, IRISA, INRIA, FR  
**Co-Chair:** Gabriela Nicolescu, Polytechnique Montréal, CA

---

### 11.1.1 SYSTEM-LEVEL EVALUATION OF CHIP-SCALE SILICON PHOTONIC NETWORKS FOR EMERGING DATA-INTENSIVE APPLICATIONS

**Speaker:** Ayse Coskun, Boston University, US  
**Authors:** Aditya Narayan, Yvain Thonnart, Pascal Vivet, Ajay Joshi and Ayse Coskun  
1Boston University, US; 2CEA-Leti, FR

**Abstract**  
Emerging data-driven applications such as graph processing applications are characterized by their excessive memory footprint and abundant parallelism, resulting in high memory bandwidth demand. As the scale of datasets for applications are reaching orders of TBs, performance limitation due to bandwidth demands is a major concern. Traditional on-chip electrical networks fail to meet such high bandwidth demands due to increased energy-per-bit or physical limitations with pin counts. Silicon photonic networks have emerged as a promising alternative to electrical interconnects, owing to their high bandwidth and low energy-per-bit communication with negligible data-dependent power. Wide-scale adoption of silicon photonics at chip level, however, is hampered by their high sensitivity to process and thermal variations, high laser power due to losses along the network, and power consumption of the electrical-optical conversion. Device-level technological innovations to mitigate these issues are promising, yet they do not consider the system-level implications of the applications running on manycore systems with photonic networks. This work aims to bridge the gap between the system-level attributes of applications with the underlying architectural and device-level characteristics of silicon photonic networks to achieve energy-efficient computing. We particularly focus on graph applications, which involve unstructured yet abundant parallel memory accesses that stress the on-chip communication networks, and develop a cross-layer framework to evaluate 2.5D systems with silicon photonic networks. We demonstrate significant energy savings through system-level management using wavelength selection policies and further evaluate architectural design choices on 2.5D systems with photonic networks.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

### 11.1.2 OSCAR: AN OPTICAL STOCHASTIC COMPUTING ACCELERATOR FOR POLYNOMIAL FUNCTIONS

**Speaker:** Sébastien Le Beux, Concordia University, CA  
**Authors:** Hassnaa El-Derhalli, Sébastien Le Beux and Sofiène Tahar, Concordia University, CA

**Abstract**  
Approximate computing allows to trade-off design energy efficiency with computing accuracy. Stochastic computing is an approximate computing technique, where numbers are represented as probabilities using stochastic bit streams. The serial computation of the bit streams leads to reduced hardware complexity but induces high latency, which is the main limitation of the approach. Silicon photonics has the potential to overcome the processing latency drawback thanks to high-speed propagation of signals and large bandwidth. However, the implementation of stochastic computing architectures using integrated optics involves high static energy that calls for adaptable architectures able to meet application-specific requirements. In this paper, we propose a reconfigurable optical accelerator allowing online adaptation of computing accuracy and energy efficiency according to the application requirements. The architecture can be configured to execute i) 4th order function for high accuracy processing or ii) 2nd order function for high-energy efficiency purposes. Evaluations are carried out using image processing Gamma correction function. Compared to a static architecture for which accuracy is defined at design time, the proposed architecture leads to 36.8% energy overhead but increases the range of reachable accuracy by 65%.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

### 11.1.3 POPSTAR: A ROBUST MODULAR OPTICAL NOC ARCHITECTURE FOR CHIPLET-BASED 3D INTEGRATED SYSTEMS

**Speaker:** Yvain Thonnart, CEA-Leti, FR  
**Authors:** Yvain Thonnart, Stéphane Bernabe, Jean Charbonnier, César Fuget Totolero, Pierre Tissier, Benoit Charbonnier, Stephane Malhoute, Damien Saint-Patrice, Myriam Assous, Aditya Narayan, Ayse Coskun, Denis Dutot and Pascal Vivet  
1CEA-Leti, FR; 2Boston University, US

**Abstract**  
Silicon photonics technology is now gaining maturity with increasing levels of design complexity from devices to large photonic integrated circuits. Close integration of control electronics with 3D assembly of photonics and CHOS open the way to high-performance computing architectures partitioned in chiplets connected by optical NoC on silicon photonic interspersers. In this paper, we give an overview of our works on optical links and NoC for manycore systems, from low-level control of photonic devices to high-level system optimization of the optical communications. We detail the POPSTAR architecture (Processors On Photonic Silicon Interposer Terascale ARchitecture) with electro-optical interface chiplets, the corresponding nested spiral topology for single-writer multiple-reader links and the associated control electronics, in charge of high-speed drivers, thermal stabilization and handling of the protocol stack, from data integrity to flow-control, routing and arbitration of the optical communications. The strengths and opportunities for this architecture will be discussed, with a shift in system & implementation constraints with respect to previous optical NoC proposals, and new challenges to be addressed.

Download Paper (PDF; Only available from the DATE venue WiFi)

**Date:** Thursday 12 March 2020  
**Time:** 14:00 - 15:30  
**Location / Room:** Chamrousse

**Chair:** Nikos Aréchiga, Toyota Research Institute, US  
**Co-Chair:** Jyotirmoy V. Deshmukh, University of Southern California, US

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00</td>
<td>11.2.1</td>
<td><strong>TRUSTWORTHY AUTONOMY: BEHAVIOR PREDICTION AND VALIDATION</strong></td>
<td>Katherine Driggs-Campbell, University of Illinois Urbana Champaign, US</td>
</tr>
<tr>
<td>14:30</td>
<td>11.2.2</td>
<td><strong>ON INFUSING LOGICAL REASONING INTO ROBOT LEARNING</strong></td>
<td>Marco Pavone, Stanford University, US</td>
</tr>
<tr>
<td>15:00</td>
<td>11.2.3</td>
<td><strong>FORMALLY-SPECIFIABLE AGENT BEHAVIOR MODELS FOR AUTONOMOUS VEHICLE TEST GENERATION</strong></td>
<td>Jonathan DeCastro, Toyota Research Institute, US</td>
</tr>
<tr>
<td>15:30</td>
<td></td>
<td><strong>End of session</strong></td>
<td></td>
</tr>
</tbody>
</table>

### 11.3 Special Session: Emerging Neural Algorithms and Their Impact on Hardware

**Date:** Thursday 12 March 2020  
**Time:** 14:00 - 15:30  
**Location / Room:** Autrans

**Chair:** Ian O’Connor, Ecole Centrale de Lyon, FR  
**Co-Chair:** Michael Niemier, University of Notre Dame, US

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00</td>
<td>11.3.1</td>
<td><strong>ANALOG RESISTIVE CROSSBAR ARRAYS FOR NEURAL NETWORK ACCELERATION</strong></td>
<td>Martin Frank, IBM, US</td>
</tr>
<tr>
<td>14:30</td>
<td>11.3.2</td>
<td><strong>IN-MEMORY COMPUTING FOR MEMORY AUGMENTED NEURAL NETWORKS</strong></td>
<td>X. Sharon Hu (^1) and Anand Raghunathan (^2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(^1)University of Notre Dame, US; (^2)Purdue University, US</td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td>11.3.3</td>
<td><strong>HARDWARE CHALLENGES FOR NEURAL RECOMMENDATION SYSTEMS</strong></td>
<td>Udit Gupta, Harvard University, US</td>
</tr>
<tr>
<td>15:30</td>
<td></td>
<td><strong>End of session</strong></td>
<td></td>
</tr>
</tbody>
</table>

### 11.4 Reliable in-memory computing

**Date:** Thursday 12 March 2020  
**Time:** 14:00 - 15:30  
**Location / Room:** Stendhal

**Chair:** Jean-Philippe Noel, CEA-Leti, FR  
**Co-Chair:** Kwatinisky Shahar, Technion, IL

This session deals with work on the reliability of computing in memories. This includes new design techniques to improve CNN computing in ReRAM going through the co-optimization between device and algorithm to improve the reliability of ReRAM-based Graph Processing. Moreover, this session also deals with work on the improvement of reliability of well-established STT-MRAM and PCM. Finally, early works presenting stochastic computing and disruptive image processing techniques based on memristor are also discussed.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Label</td>
<td>Presentation Title</td>
<td>Speaker</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>------------------------------------------------------------------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>14:00</td>
<td>11.4.1</td>
<td>REBOC: ACCELERATING BLOCK-CIRCULANT NEURAL NETWORKS IN ReRAM</td>
<td>Yitu Wang, Fudan University, CN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:30</td>
<td>11.4.2</td>
<td>GRAPHPRISM: A JOINT DEVICE-ALGORITHM RELIABILITY ANALYSIS FOR ReRAM-BASED GRAPH PROCESSING</td>
<td>Chin-Fu Nien, Academia Sinica, TW</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td>11.4.3</td>
<td>STAIR: HIGH RELIABLE STT-MRAM AWARE MULTI-LEVEL I/O CACHE ARCHITECTURE BY ADAPTIVE ECC ALLOCATION</td>
<td>Hossein Asadi, Sharif University of Technology, IR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:15</td>
<td>11.4.4</td>
<td>EFFECTIVE WRITE DISTURBANCE MITIGATION ENCODING SCHEME FOR HIGH-DENSITY PCM</td>
<td>Muhammad Imran, Sungkyunkwan University, KR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Computational Restructuring: Rethinking Image Processing Using Memristor Crossbar Arrays

Speaker: Rickard Ewetz, University of Central Florida, US

Authors: Baogang Zhang, Necati Uysal and Rickard Ewetz, University of Central Florida, US

Abstract

Image processing is a core operation performed on billions of sensor-devices in the Internet of Things (IoT). Emerging memristor crossbar arrays (MCAs) promise to perform matrix-vector multiplication (MVM) with extremely small energy-delay product, which is the dominating computation within the two-dimensional Discrete Cosine Transform (2D DCT). Earlier studies have directly mapped the digital implementation to MCA based hardware. The drawback is that the series computation is vulnerable to errors. Moreover, the implementation requires the use of large image block sizes, which is known to degrade the image quality. In this paper, we propose to restructure the 2D DCT into an equivalent single linear transformation (or MVM operation). The reconstruction eliminates the series computation and reduces the processed block sizes from N x N to √N x √N. Consequently, both the robustness to errors and the image quality is improved. Moreover, the latency, power, and area is reduced with 2X while eliminating the storage of intermediate data, and the power and area can be further reduced with up to 62% and 74% using frequency spectrum optimization.

Download Paper (PDF; Only available from the DATE venue WiFi)

SCRIMP: A Generalized Stochastic Computing Acceleration Architecture Using ReRAM In-Memory Processing

Speaker: Saransh Gupta, University of California, San Diego, US

Authors: Saransh Gupta1, Mohsen Imani1, Joonsop Sim1, Andrew Huang1, Fan Wu1, M. Hassan Najafi2 and Tajana Rosing1

1University of California, San Diego, US; 2University of Louisiana, US

Abstract

Stochastic computing (SC) reduces the complexity of computation by representing numbers with long independent bit-streams. However, increasing performance in SC comes with an increase in area and loss in accuracy. Processing in memory (PIM) with non-volatile memories (NVMs) computes data in-place, while having high memory density and supporting bit-parallel operations with low energy. In this paper, we propose SCRIMP for stochastic computing acceleration with resistive RAM (ReRAM) in-memory processing, which enables SC in memory. SCRIMP can be used for a wide range of applications. It supports all SC encodings and operations in memory. It maximizes the performance and energy efficiency of implementing SC by introducing novel in-memory parallel stochastic number generation and efficient implication-based logic in memory. To show the efficiency of our stochastic hardware, we implement image processing on the proposed hardware.

Download Paper (PDF; Only available from the DATE venue WiFi)

11.5 Compile time and virtualization support for embedded system design

Date: Thursday 12 March 2020
Time: 14:00 - 15:30
Location / Room: Bayard

Chair: Nicola Bombieri, Università di Verona, IT
Co-Chair: Rodolfo Pellizzoni, University of Waterloo, CA

The session leverages compiler support and novel architectural features, such as virtualization extensions and emerging memory structures, to optimize the design flow of modern embedded systems.

Unified Thread- and Data-Mapping for Multi-Threaded Multi-Phase Applications on SPM Many-Cores

Speaker: Anuj Pathania, National University of Singapore, SG

Authors: Vanchinthan Venkataramani, Anuj Pathania and Tulika Mitra, National University of Singapore, SG

Abstract

Scratchpad Memories (SPMs) are more scalable than caches as they offer better performance with lower power and area overheads. This scalability advocates their suitability as on-chip memory in many-cores. However, SPM many-cores delegate the responsibility of thread- and data-mapping to the software. The mapping is especially challenging in the case of multi-threaded multi-phase applications. Threads from these applications exhibit both inter- and intra-phase data-sharing patterns. These patterns intrinsically intertwine thread- and data-mapping across phases. The accompanying qualitative motivation is the key to extract application performance on SPM many-cores. State-of-the-art framework for SPM many-cores performs thread- and data-mapping independently. Furthermore, it can only operate with single-phase multi-threaded applications. We are the first to propose in this work, a unified thread- and data-mapping framework for NoC-based SPM many-cores when executing multi-threaded multi-phase applications. Experimental evaluations show, on average, 1.36x performance improvement compared to the state-of-the-art framework for multi-threaded multi-phase applications.

Download Paper (PDF; Only available from the DATE venue WiFi)

Generalized Data Placement Strategies for Racetrack Memories

Speaker: Asif Ali Khan, TU Dresden, DE

Authors: Asif Ali Khan, Andres Goens, Fazal Hameed and Jeronimo Castrillon, TU Dresden, DE

Abstract

Ultra-dense non-volatile racetrack memories (RTMs) have been investigated at various levels in the memory hierarchy for improved performance and reduced energy consumption. However, the innate shift operations in RTMs hinder their applicability to replace low-latency on-chip memories. Recent research has demonstrated that intelligent placement of memory objects in RTMs can significantly reduce the amount of shifts with no hardware overhead, albeit for specific system setups. However, existing placement strategies may lead to sub-optimal performance when applied to different architectures. In this paper we look at generalized data placement mechanisms that improve upon existing ones by taking into account the underlying memory architecture and the timing and liveliness information of memory objects. We propose a novel heuristic and a formulation using genetic algorithms that optimize key performance parameters. We show that, on average, our generalized approach improves the number of shifts, performance and energy consumption by 4.3x, 46% and 55% respectively compared to the state-of-the-art.

Download Paper (PDF; Only available from the DATE venue WiFi)
### 11.6 Aging: estimation and mitigation

**Date:** Thursday 12 March 2020  
**Time:** 14:00 – 15:30  
**Location / Room:** Lesdiguières

**Chair:**  
Arnaud Virazel, Université de Montpellier / LIRMM, FR

**Co-Chair:**  
Lorena Anghel, University Grenoble-Alpes, FR

This session shares improvements in aging calculations of emerging technologies and how to take these reliability aspects into account during power grid design and floorplanning of FPGAs.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
</table>
| 15:30 | IPS-8 | ARM-ON-ARM: LEVERAGING VIRTUALIZATION EXTENSIONS FOR FAST VIRTUAL PLATFORMS | Lukas Jünger, RWTH Aachen University, DE  
Lukas Jünger¹, Jan Luca Malte Bölke², Stephan Tobies², Rainer Leupers¹ and Andreas Hoffmann ²  
¹RWTH Aachen University, DE; ²Synopsys GmbH, DE |

**Abstract**  
Virtual Platforms (VPs) are an essential enabling technology in the System-on-a-Chip (SoC) development cycle. They are used for early software development and hardware/software codesign. However, since virtual prototyping is limited by simulation performance, improving the simulation speed of VPs has been an active research topic for years. Different strategies have been proposed, such as fast instruction set simulation using Dynamic Binary Translation (DBT). But even fast simulators do not reach native execution speed. They do however allow executing rich Operating System (OS) kernels, which is typically infeasible when another OS is already running. Executing multiple OSs on shared physical hardware is typically accomplished by using virtualization, which has a long history on x86 hardware. It enables encapsulated, native code execution on the host processor and has been extensively used in data centers, where many users share hardware resources. When it comes to embedded systems, virtualization has been made available recently. For ARM processors, virtualization was introduced with the ARM Virtualization Extensions for the ARMv7 architecture. Since virtualization allows native guest code execution, near-native execution speeds can be reached. In this work we present a VP containing a novel ARMv8 SystemC Transaction Level Modeling (TLM) compatible processor model. The model leverages the ARM Virtualization Extensions (VE) via the Linux Kernel-based Virtual Machine (KVM) to execute the target software natively on an ARMv8 host. To enable the integration of the processor model into a loosely-timed VP, we developed an accurate instruction counting mechanism using the ARM Performance Monitors Extension (PMU). The requirements for integrating the processor model into a VP and the integration process are detailed in this work. Our evaluations show that speedups of up to 2.57x over state-of-the-art DBT-based simulator can be achieved using our processor model on ARMv8 hardware. |

**Download Paper (PDF; Only available from the DATE venue WiFi)**

| 15:30 | IPS-9 | TDO-CIM: TRANSPARENT DETECTION AND OFFLOADING FOR COMPUTATION IN-MEMORY | Lorenzo Chelini, Eindhoven University of Technology, NL  
Kanishkan Vadivel¹, Lorenzo Chelini², Ali BanaGozar², Gagandeep Singh², Stefano Corda², Roel Jordans¹ and Henk Corporaal¹ |

**Abstract**  
Computation in-memory is a promising non-von Neumann approach aiming at completely diminishing the data transfer to and from the memory subsystem. Although a lot of architectures have been proposed, compiler support for such architectures is still lagging behind. In this paper, we close this gap by proposing an end-to-end compilation flow for in-memory computing based on the LLVM compiler infrastructure. Starting from sequential code, our approach automatically detects, optimizes, and offloads kernels suitable for in-memory acceleration. We demonstrate our compiler tool-flow on the PolyBench/C benchmark suite and evaluate the benefits of our proposed in-memory architecture simulated in Gem5 by comparing it with a state-of-the-art von Neumann architecture. |

**Download Paper (PDF; Only available from the DATE venue WiFi)**

| 15:30 | IPS-9 | BACKFLOW: BACKWARD EDGE CONTROL FLOW ENFORCEMENT FOR LOW END ARM MICROCONTROLLERS | Cyril Bresch, LCIS, FR  
Cyril Bresch¹, David Hély² and Roman Lysecky³ |

**Abstract**  
This paper presents BackFlow, a compiler-based toolchain that enforces indirect backward edge control flow integrity for low-end ARM Cortex-M microprocessors. BackFlow is implemented within the Clang/LLVM compiler and supports the ARM instruction set and its subset Thumb. The control flow integrity generated by the compiler relies on a bitmap, where each set bit indicates a valid pointer destination. The efficiency of the framework is benchmarked using an STM32 NUCLEO F446RE microcontroller. The obtained results show that the control flow integrity solution incurs an execution time overhead ranging from 1.5 to 4.5%. |

**Download Paper (PDF; Only available from the DATE venue WiFi)**

End of session
This is the first work that investigates the impact of Negative Bias Temperature Instability (NBTI) on the Self-Heating (SH) phenomena in Silicon Nanowire Field-Effect Transistors (SiNW-FETs). We investigate the individual as well as joint impact of NBTI and SH on SiNW-FETs and demonstrate that NBTI-induced traps mitigate SH effects due to reduced current densities. Our Technology CAD (TCAD)-based SiNW-FET device is calibrated against experimental data. It accounts for thermodynamic and hydrodynamic effects in 3-D nano structures for accurate modeling of carrier transport mechanisms. Our analysis focuses on how lattice temperature, thermal resistance and thermal capacitances of SiNW-FETs are affected due to NBTI, demonstrating that accurate self-heating modeling necessitates considering the effects that NBTI aging has over time. Hence, NBTI and SH effects need to be jointly and not individually modeled. Our evaluation shows that an individual modeling of NBTI and SH effects leads to a noticeable overestimation of the overall induced delay increase in circuits due to the impact of NBTI traps on SH mitigation. Hence, it is necessary to model NBTI and SH effects jointly in order to estimate efficient (i.e. small, yet sufficient) timing guardbands that protect circuits against timing violations, which will occur at runtime due to delay increases induced by aging and self-heating.

Abstract

Daniel Kraak, MIT

IMPACT OF NBTI AGING ON SELF-HEATING IN NANOWIRE FET

Speaker:

Hussam Arnrouch, Karlsruhe Institute of Technology, DE

Authors:

Om Prakash1, Hussam Arnrouch 1, Sanjeev Kumar Manhas2 and Joerg Henkel 1

1Karlsruhe Institute of Technology, DE; 2IIT Roorkee, IN

11.6.2 POWERPLANNINGDL: RELIABILITY-AWARE FRAMEWORK FOR ON-CHIP POWER GRID DESIGN USING DEEP LEARNING

Speaker:

Sukanta Dey, IIT Guwahati, IN

Authors:

Sukanta Dey, Sukumar Nandi and Gaurav Trivedi, IIT Guwahati, IN

Abstract:

With the increase in complexity of chip designs, VLSI physical design has become a time-consuming task, which is an iterative design process. Power planning is that part of the floorplanning in VLSI physical design where power grid networks are designed in order to provide adequate power to all the underlying functional blocks. Power planning also requires multiple iterative steps to create the power grid network while satisfying the allowed worst-case IR drop and Electromigration (EM) margin. For the first time, this paper introduces Deep learning (DL)-based framework to approximately predict the initial design of the power grid network, considering different reliability constraints. The proposed framework reduces many iterative design steps and speeds up the total design cycle. Neural Network-based multi-target regression technique is used to create the DL model. Feature extraction is done, and training dataset is generated from the floorplans of some of the power grid designs extracted from IBM processor. The DL model is trained using the generated dataset. The proposed DL-based framework is validated using a new set of power grid specifications (obtained by perturbing the designs used in the training phase). The results show that the predicted power grid design is closer to the original design with minimal prediction error (~2%). The proposed DL-based approach also reduces the design cycle time significantly with a speedup of ~6X for standard power grid benchmarks.

Download Paper (PDF; Only available from the DATE venue WiFi)

11.6.3 AN EFFICIENT MILP-BASED AGING-AWARE FLOORPLANNER FOR MULTI-CONTEXT COARSE-GRAINED RUNTIME RECONFIGURABLE FPGAS

Speaker:

Carl Sechen, University of Texas at Dallas, US

Authors:

Bo Hu, Mustafa Shihab, Yiorgos Makris, Benjamin Carnin Schaefer and Carl Sechen, University of Texas at Dallas, US

Abstract:

Shrinking transistor sizes are jeopardizing the reliability of runtime reconfigurable Field Programmable Gate Arrays (FPGAs), making them increasingly sensitive to aging effects such as Negative Bias Temperature Instability (NBTI). This paper introduces a reliability-aware floorplanner which is tailored to multi-context, coarse-grained, runtime reconfigurable architectures (CGRRAs) and seeks to extend their Mean Time to Failure (MTTF) by balancing the usage of processing elements (PEs). The proposed method is based on a Mixed Integer Linear Programming (MILP) formulation, the solution to which produces appropriately-balanced mappings of workload to PEs on the reconfigurable fabric, thereby mitigating aging-induced lifetime degradation. Results demonstrate that, as compared to the default reliability-unaware floorplanning solutions, the proposed method achieves an average MTTF increase of 2.5X without introducing any performance degradation.

Download Paper (PDF; Only available from the DATE venue WiFi)

11.7 System Level Security
The session focuses on topics of system-level security, especially related to authentication. The papers span topics of memory authentication and group-of-users authentication, with a focus on IoT applications.

**Time** | **Label** | **Presentation Title** | **Authors**
--- | --- | --- | ---
14:00 | 11.7.1 | **AMSA: ADAPTIVE MERKLE SIGNATURE ARCHITECTURE** | Emanuel Regnath, TU Munich, DE; Emanuel Regnath and Sebastian Steinhorst, TU Munich, DE

**Abstract**
Hash-based signatures (HBS) are promising candidates for quantum-secure signatures on embedded IoT devices because they only use fast integer math, are well understood, produce small public keys, and offer many design parameters. However, HBS can only sign a limited amount of messages and produce - similar to most post-quantum schemes - large signatures of several kilo bytes. In this paper, we explore possibilities to reduce the size of the signatures by 1. improving the Winternitz One-Time Signature with a more efficient encoding and 2. ofloading auxiliary data to a gateway. We show that for similar security and performance, our approach produces 2.6 % smaller signatures in general and up to 17.3 % smaller signatures for the sender compared to the related approaches LMS and XMSS. Furthermore, our open-source implementation allows a wider set of parameters that allows to tailor the scheme to the available resources of an embedded device, which is an important factor to overcome the security challenges in IoT.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

14:30 | 11.7.2 | **DISSECT: DYNAMIC SKEW-AND-SPLIT TREE FOR MEMORY AUTHENTICATION** | Lam Siew-Kei, Nanyang Technological University, SG; Saru Vigi, Rohan Juneya, Siew Kei Lam, Qualcomm, IN

**Abstract**
Memory integrity trees are widely-used to protect external memories in embedded systems against replay, splicing and spoofing attacks. However, existing methods often result in high-performance overhead that is proportional to the height of the tree. Reducing the height of the integrity tree by increasing its arity, however, leads to frequent overflowing of the counters that are used for encryption in the tree. We will show that increasing the tree arity of a widely-used integrity tree from 2 to 8 can result in over 200% increase in memory authentication overhead for some benchmark applications, despite the reduction in tree height. In this paper, we propose DISSECT, a memory authentication framework which utilizes a dynamic memory integrity tree that can adapt to the memory access patterns of the application by progressively adjusting the tree height and arity in order to significantly reduce performance overhead. This is achieved by 1) initializing an integrity tree structure with the largest arity possible considering the performance impact due to counter overflow, 2) dynamically skewing the tree such that the more frequently accessed memory locations are positioned closer to the tree root (overcomes the tree height problem), and 3) dynamically splitting the tree at nodes with counters that are about to overflow (overcomes the counter overflow problem). Experimental results undertaken using Multi2Sim on benchmarks from SPEC-CPU2006, SPLASH-2, and PARSEC demonstrate the performance benefits of our proposed memory integrity tree.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

15:00 | 11.7.3 | **DESIGN-FLOW METHODOLOGY FOR SECURE GROUP ANONYMOUS AUTHENTICATION** | Rashmi Agrawal, Boston University, US; Rashmi Agrawal, Lake Bu, Eliakin del Rosario and Michel Kinsky, Boston University, US; 2Draper Lab, US

**Abstract**
In heterogeneous distributed systems, the computing devices and software components often come from different providers and have different security, trust, and privacy levels. In many of the systems, the need frequently arises to (i) control the access to services and resources granted to the individual devices or components in a context-aware manner, and (ii) establish and enforce data sharing policies that preserve the privacy of the critical information on end-users. In essence, the need is to simultaneously authenticate and anonymize an entity or device, two seemingly contradictory goals. The design challenge is further complicated by potential security problems such as man-in-the-middle attacks, hijacked devices, and counterfeits. In this work, we present a system design flow for a trustworthy group anonymous authentication protocol (GAAP), which not only fulfills the desired functionality for authentication and privacy, but also provides strong security guarantees.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

15:30 | IPS-12 | **BLOCKCHAIN TECHNOLOGY ENABLED PAY PER USE LICENSING APPROACH FOR HARDWARE IPS** | Krishnendu Guha, University of Calcutta, IN

**Abstract**
The present era is witnessing a reuse of hardware IPs to reduce cost. As trustworthiness is an essential factor, designers prefer to use hardware IPs which performed effectively in the past, but at the same time, are still active and did not age. In such scenarios, pay per use licensing schemes suit best for both producers and users. Existing pay per use licensing mechanisms consider a centralized third party, which may not be trustworthy. Hence, we seek refuge to blockchain technology to eradicate such third parties and facilitate a transparent and automated pay per use licensing mechanism. A blockchain is a distributed public ledger whose records are added based on peer review and majority consensus of its participants, that cannot be tampered or modified later. Smart contracts are deployed to facilitate the mechanism. Even dynamic pricing of the hardware IPs based on the factors of trustworthiness and aging have been focused in this work, which are not associated in existing literature. Security analysis of the proposed mechanism has been provided. Performance evaluation is carried based on the gas usage of Ethereum Solidity test environment, along with cost analysis based on lifetime and related user ratings.

**Download Paper (PDF; Only available from the DATE venue WiFi)**

15:30 | End of session
State-of-the-art electronic design allows the integration of complex electronic systems comprising thousands of high-level functions on a single chip. This has become possible and feasible because of the combination of atomic-scale semiconductor technology allowing VLSI of billions of transistors, and EDA tools that can handle their useful application and integration by following strictly hierarchical design methodology. This results in many layers of abstraction within a system that makes it implementable, verifiable and, ultimately, explainable. However, while many layers of abstraction maximise the likelihood of a system to function correctly, this can prevent a design from making full use of the capabilities of current technology. Making systems brittle at a time where NoC- and SoC-based technology. Making are the only way to increase compute capabilities as clock speed limits are reached, devices are affected by variability and ageing, and heat-dissipation limits impose "dark silicon" constraints. Design challenges of electronic systems are no longer driven by making designs smaller but by creating systems that are ultra-low power, resilient and autonomous in their adaptation to anomalies including faults, timing violations and performance degradation. This gives rise to the idea of self-aware hardware, capable of adaptive behaviours or features taking inspiration from, e.g., biological systems, learning algorithms, factory processes. The challenge is to adopt and implement these concepts while achieving a "next- generation" kind of electronic system which is considered at least as useful and trustworthy as its "classical" counterpart—plus additional essential features for future system design and operation. The goal of this Special Session is to present research from world-leading experts addressing state-of-the-art techniques and devices demonstrating the efficacy of concepts of self-awareness, adaptivity and bio-inspiration in the context of real-world hardware systems and applications with a focus on autonomous resource management at runtime, robustness and performance, and new computing concepts in embedded hardware systems."
UB11.1 ATECES: AUTOMATED TESTING THE ENERGY CONSUMPTION OF EMBEDDED SYSTEMS
Authors: Eduard Enou, Mälardalen University, SE
Abstract
The demonstrator will focus on automatically generating test suites by selecting test cases using random test generation and mutation testing as a solution for improving the efficiency and effectiveness of testing. Specifically, we generate and select test cases based on the concept of energy-aware mutants, small syntactic modifications in the system architecture, intended to mimic real energy faults. Test cases that can distinguish a certain behavior from its mutations are sensitive to changes, and hence considered to be good at detecting faults. We applied this method on a brake by wire system and our results suggest that an approach that selects test cases showing diverse energy-consumption can increase the fault detection ability. This kind of results should motivate both academia and industry to investigate the use of automatic test generation for energy consumption.
More information ...

UB11.2 BROOK SC: HIGH-LEVEL PROGRAMMING-FRIENDLY PROGRAMMING FOR GPU-POWERED SAFETY CRITICAL SYSTEMS
Authors: Marc Benito, Matina Maria Trompouki and Leonidas Kosmidis, BSC / UPC, ES
Abstract
Graphics processing units (GPUs) can provide the increased performance required in future critical systems, i.e. automotive and avionics. However, their programming models, e.g. CUDA or OpenCL, cannot be used in such systems as they violate safety critical programming guidelines. Brook SC (https://github.com/lkosmid/brook) was developed in UPC/BSC to allow safety-critical applications to be programmed in a CUDA-like GPU language, Brook, which enables the certification while increasing productivity. In our demo, an avionics application running on a realistic safety critical GPU software stack and hardware is show cased. In this Bachelor’s thesis project, which was awarded a 2019 HiPEAC Technology Transfer Award, an Airbus prototype application performing general-purpose computations with a safety-critical graphics API was ported to Brook SC in record time, achieving an order of magnitude reduction in the lines of code to implement the same functionality without performance penalty.
More information ...

UB11.3 DISTRIBUTING TIME-SENSITIVE APPLICATIONS ON EDGE COMPUTING ENVIRONMENTS
Authors: Eudald Sabaté Creixell, Unai Perez Mendizabal, Elii Kartsaki, Maria A. Serrano Gracia and Eduardo Quiliones Moreno
Abstract
The proposed demonstration aims to showcase the capabilities of a task-based distributed programming framework for the execution of real-time applications in edge computing scenarios, in the context of smart cities. Edge computing shifts the computation close to the data source, alleviating the pressure on the cloud and reducing application response times. However, the development and deployment of distributed real-time applications is complex, due to the heterogeneous and dynamic edge environment where resources may not always be available. To address these challenges, our demo employs COMPSs, a highly portable and infrastructure-agnostic programming model, to efficiently distribute time-sensitive applications across the compute continuum. We will exhibit how COMPSs distributes the workload on different edge devices (e.g., NVIDIA GPUs and a Raspberry Pi), and how COMPSs re-adapts this distribution upon the availability (connection or disconnection) of devices.
More information ...

UB11.4 DL PUF ENU: DEEP LEARNING BASED PHYSICALLY UNCLONABLE FUNCTION ENROLLMENT AND AUTHENTICATION
Authors: Amir Alipour, David Hely, Vincent Beroille and Giorgio Di Natale
Abstract
Physically Unclonable Functions (PUFs) have been addressed nowadays as a potential solution to improve the security in authentication and encryption process in Cyber Physical Systems. The research on PUF is actively growing due to its potential of being secure, easily implementable and expandable, using considerably less energy. To use PUF in common, the low level device Hardware Variation is captured per unit for device enrollment into a format called Challenge-Response Pair (CRP), and recaptured after device is deployed, and compared with the original for authentication. These enrollment + comparison functions can vary and be more data demanding for applications that demand robustness, and resilience to noise. In this demonstration, our aim is to show the potential of using Deep Learning for enrollment and authentication of PUF CRPs. Most importantly, during this demonstration, we will show how this method can save time and storage compared to other classical methods.
More information ...

UB11.5 LAGARTO: FIRST SILICON RISC-V ACADEMIC PROCESSOR DEVELOPED IN SPAIN
Abstract
Open hardware is a possibility that has emerged in recent years and has the potential to be as disruptive as Linux was once, an open source software paradigm. If Linux managed to lessen the dependence of users in large companies providing software and software applications, it is envisioned that hardware based on ISA open source can do the same in their own field. In the Lagarto tapeout four research institutions were involved: Centro de Investigación en Computación of the Mexican IPN, Centro Nacional de Microelectrónica de the CSIC, Universitat Politècnica de Catalunya (UPC) and Barcelona Supercomputing Center (BSC). As a result, many bachelor, master and PhD students had the chance to achieve real-world experience with ASIC design and achieve a functional SoC. In the booth, you will find a live demo of the first ASIC and prototypes running on FPGA of the next versions of the SoC and core.
More information ...

UB11.6 SRSN: SECURE RECONFIGURABLE TEST NETWORK
Authors: Vincent Reynaud, Emanuele Valez, Paolo Maistri, Regis Leveugle, Marie-Lise Flottes, Sophie Dupuis, Bruno Rouzeure and Giorgio Di Natale
Abstract
The critical importance of testability for electronic devices led to the development of IEEE test standards. These methods, if not protected, offer a security backdoor to attackers. This demonstrator illustrates a state-of-the-art solution that prevents unauthorized usage of the test infrastructure based on the IEEE 1687 standard and implemented on an FPGA target.
More information ...
**Abstract**

In order to spread asynchronous circuit design to a large community of designers, High-Level Synthesis (HLS) is probably a good choice because it requires limited design technical skills. HLS usually provides an RTL description, which includes a data-path and a control-path. The desynchronization process is only applied to the control-path, which is a Finite State Machine (FSM). This method is sufficient to make asynchronous the circuit. Indeed, data are processed step by step in the pipeline stages, thanks to the desynchronized FSM. Thus, the data-path computation time is no longer related to the clock period but rather to the average time for processing data into the pipeline. This tends to improve speed when the pipeline stages are not well-balanced. Moreover, our approach helps to quickly designing data-driven circuits while maintaining a reasonable cost, a similar area and a short time-to-market. However, we have to be aware of the overheads introduced by the approach, which can be significant.

**More information ...**
SCRIMP: A GENERAL STOCHASTIC COMPUTING ACCELERATION ARCHITECTURE USING RERAM IN-MEMORY PROCESSING

Speaker: Saransh Gupta, University of California, San Diego, US

Authors: Saransh Gupta1, Mohsen Imani1, Joonseop Sim1, Andrew Huang1, Fan Wu1, M. Hassan Najafi2 and Tajana Rosing1
1University of California, San Diego, US; 2University of Louisiana, US

Abstract: Stochastic computing (SC) reduces the complexity of computation by representing numbers with long independent bit-streams. However, increasing performance in SC comes with an increase in area and loss in accuracy. Processing in memory (PIM) with non-volatile memories (NVMs) computes data in-place, while having high memory density and supporting bit-parallel operations with low energy. In this paper, we propose SCRIMP for stochastic computing acceleration with resistive RAM (ReRAM) in-memory processing, which enables SC in memory. SCRIMP can be used for a wide range of applications. It supports all SC encodings and operations in memory. It maximizes the performance and energy efficiency of implementing SC by introducing novel in-memory parallel stochastic number generation and efficient implication-based logic in memory. To show the efficiency of our stochastic architecture, we implement image processing on the proposed hardware.

Download Paper (PDF; Only available from the DATE venue WiFi)
**IPS-8**

**Title:** TD-O-CIM: TRANSPARENT DETECTION AND OFFLOADING FOR COMPUTATION IN-MEMORY

**Speaker:** Lorenzo Chelini, Eindhoven University of Technology, NL

**Authors:** Kanishkan Vadivel1, Lorenzo Chelini2, Ali BanaGozar1, Gagandeep Singh2, Stefano Corda2, Roel Jordans1 and Henk Corporaal1

1Eindhoven University of Technology, NL; 2IBM Research, CH

**Abstract**

Computation in-memory is a promising non-von Neumann approach aiming at completely diminishing the data transfer to and from the memory subsystem. Although a lot of architectures have been proposed, compiler support for such architectures is still lagging behind. In this paper, we close this gap by proposing an end-to-end compilation flow for in-memory computing based on the LLVM compiler infrastructure. Starting from sequential code, our approach automatically detects, optimizes, and offloads kernels suitable for in-memory acceleration. We demonstrate our compiler tool-flow on the PolyBench/C benchmark suite and evaluate the benefits of our proposed in-memory architecture simulated in Gem5 by comparing it with a state-of-the-art von Neumann architecture.

Download Paper (PDF; Only available from the DATE venue WiFi)

**IPS-9**

**Title:** BACKFLOW: BACKWARD EDGE CONTROL FLOW ENFORCEMENT FOR LOW END ARM MICROCONTROLLERS

**Speaker:** Cyril Bresch, LCIS, FR

**Authors:** Cyril Bresch1, David Hély2 and Roman Lysecky1

1LCIS, FR; 2LCIS - Grenoble INP, FR; 3University of Arizona, US

**Abstract**

This paper presents BackFlow, a compiler-based toolchain that enforces indirect backward edge control flow integrity for low-end ARM Cortex-M microprocessors. BackFlow is implemented within the Clang/LLVM compiler and supports the ARM instruction set and its subset Thumb. The control flow integrity generated by the compiler relies on a bitmap, where each set bit indicates a valid pointer destination. The efficiency of the framework is benchmarked using an STM32 NUCLEO F446RE microcontroller. The obtained results show that the control flow integrity solution incurs an execution time overhead ranging from 1.5 to 4.5%.

Download Paper (PDF; Only available from the DATE venue WiFi)

**IPS-10**

**Title:** DELAY SENSITIVITY POLYNOMIALS BASED DESIGN-DEPENDENT PERFORMANCE MONITORS FOR WIDE OPERATING RANGES

**Speaker:** Ruikai Shi, Chinese Academy of Sciences, CN

**Authors:** Ruikai Shi1, Liang Yang2 and Hao Wang2

1Chinese Academy of Sciences / University of Chinese Academy of Sciences, CN; 2Loongson Technology Corporation Ltd., CN

**Abstract**

The downsizing of CMOS technology makes circuit performance more sensitive to on-chip parameter variations. Previous proposed design-dependent ring oscillator (DDRO) method provides an efficient way to monitor circuit performance at runtime. However, the linear delay sensitivity expression may be inadequate, especially in a wide range of operating conditions. To overcome it, a new design-dependent performance monitor (DDPM) method is proposed in this work, which formulates the delay sensitivity as high-order polynomials, makes it possible to accurately track the nonlinear timing behavior for wide operating ranges. A 28nm technology is used for design evaluation, and quite a low error rate is achieved in circuit performance monitoring comparison.

Download Paper (PDF; Only available from the DATE venue WiFi)

**IPS-11**

**Title:** MITIGATION OF SENSE AMPLIFIER DEGRADATION USING SKEWED DESIGN

**Speaker:** Daniel Kraak, TU Delft, NL

**Authors:** Daniel Kraak1, Mottaqiallah Taouil1, Said Hamdioui1, Pieter Weckx2, Stefan Cosemans2 and Francky Catthoor2

1TU Delft, NL; 2IMEC, BE

**Abstract**

Designers typically add design margins to semiconductor memories to compensate for aging. However, the aging impact increases with technology downsizing, leading to the need for higher margins. This results into a negative impact on area, yield, performance, and power consumption. As an alternative, mitigation schemes can be developed to reduce such impact. This paper proposes a mitigation scheme for the memory’s sense amplifier (SA); the scheme is based on creating a skew in the relative strengths of the SA’s cross-coupled inverters during design. The skew is compensated by aging due to unbalanced workloads. As a result, the impact of aging on the SA is reduced. To validate the mitigation scheme, the degradation of the sense amplifier is analyzed for several workloads. The experimental results show that the proposed mitigation scheme reduces the degradation of the sense amplifier’s critical figure-of-merit, the offset voltage, with up to 26%.

Download Paper (PDF; Only available from the DATE venue WiFi)

**IPS-12**

**Title:** BLOCKCHAIN TECHNOLOGY ENABLED PAY PER USE LICENSING APPROACH FOR HARDWARE IPS

**Speaker:** Krishnendu Guha, University of Calcutta, IN

**Authors:** Krishnendu Guha, Debasri Saha and Aman Chakrabarti, University of Calcutta, IN

**Abstract**

The present era is witnessing a reuse of hardware IPs to reduce cost. As trustworthiness is an essential factor, designers prefer to use hardware IPs which performed effectively in the past, but at the same time, are still active and did not age. In such scenarios, pay per use licensing schemes suit best for both producers and users. Existing pay per use licensing mechanisms consider a centralized third party, which may not be trustworthy. Hence, we seek refuge to blockchain technology to eradicate such third parties and facilitate a transparent and automated pay per use licensing mechanism. A blockchain is a distributed public ledger whose records are added based on peer review and majority consensus of its participants, that cannot be tampered or modified later. Smart contracts are deployed to facilitate the mechanism. Even dynamic pricing of the hardware IPs based on the factors of trustworthiness and aging have been focused in this work, which are not associated in existing literature. Security analysis of the proposed mechanism has been provided. Performance evaluation is carried based on the gas usage of Ethereum Solidity test environment, along with cost analysis based on lifetime and related user ratings.

Download Paper (PDF; Only available from the DATE venue WiFi)

### 12.1 Special Day on "Silicon Photonics": Design Automation for Photonics

**Date:** Thursday 12 March 2020
**Time:** 16:00 - 17:30
**Location / Room:** Amphithéâtre Jean Prouve

**Chair:** Dave Penkler, SCINTL Photonics, US
**Co-Chair:** Luca Ramini, Hewlett Packard Labs, US

**Time** | **Presentation Title** | **Authors**
--- | --- | ---

Download Paper (PDF; Only available from the DATE venue WiFi)
Abstract
With the ever growing complexity of high-performance computing (HPC) systems to satisfy emerging application requirements (e.g., high memory bandwidth requirement for machine learning applications), the performance bottleneck in such systems has moved from being computation-centric to be more communication-centric. Silicon photonic interconnection networks have been proposed to address the aggressive communication requirements in HPC systems, to realize higher bandwidth, lower latency, and better energy efficiency. There have been many successful efforts on developing silicon photonic devices, integrated circuits, and architectures for HPC systems. Moreover, many efforts have been made to address and mitigate the impact of different challenges (e.g., fabrication process and thermal variations) in silicon photonic interconnects. However, most of these efforts have focused only on a single design layer in the system design space (e.g., device, circuit or architecture level). Therefore, there is often a gap between what a design technique can improve in one layer, and what it might impair in another one. In this paper, we discuss the promise of cross-layer design methodologies for HPC systems integrating silicon photonic interconnects. In particular, we discuss how such cross-layer design solutions based on cooperatively designing and exchanging design objectives among different system design layers can help achieve the best possible performance when integrating silicon photonics into HPC systems.

Download Paper (PDF; Only available from the DATE venue WiFi)

17:00 12.1.3 EFFICIENT OPTICAL POWER DELIVERY SYSTEM FOR HYBRID ELECTRONIC-PHOTONIC MANYCORE PROCESSORS
Speaker: Jiang Xu, Hong Kong University of Science and Technology, HK
Authors: Shixi Chen, Jiang Xu, Xuanqi Chen, Zhifei Wang, Jun Feng, Jiaxu Zhang, Zhongyuan Tian and Xiao Li, Hong Kong University of Science and Technology, HK
Abstract
A lot of efforts have been devoted to optically enabled high-performance communication infrastructures for future manycore processors. Silicon photonic network promises high bandwidth, high energy efficiency and low latency. However, the ever-increasing design complexity results in complex optical power demands, which stress the optical power delivery and affect delivery efficiency. Facing optical power delivery challenges, we propose a Ring-based Optical Active Delivery (ROAD) system, to effectively manage and efficiently deliver optical power throughout photonic-electronic hybrid systems.

Download Paper (PDF; Only available from the DATE venue WiFi)

17:30 End of session
12.4 Approximate Computing Works! Applications & Case Studies

Date: Thursday 12 March 2020
Time: 16:00 - 17:30
Location / Room: Stendhal

Chair:
Oliver Keszeocze, Friedrich-Alexander-University Erlangen-Nuremberg (FAU), DE

Co-Chair:
Benjamin Carrion Schaefer, University of Texas at Dallas, US

Approximate computing leverages the fact that many applications are tolerant of incorrect results. This session highlights that by presenting methods and applications that optimize the trade-off between area, power and output error. At the same time it is important to ensure that the approximation approaches are scalable because complex problems are addressed. While some of these approaches completely work at the application level, others are oriented towards optimizing key subcircuits.
### 12.4.1 TOWARDS GENERIC AND SCALABLE WORD-LENGTH OPTIMIZATION

**Speaker:** Van-Phu Ha, Université de Rennes / Inria / IRISA, FR  
**Authors:** Van-Phu Ha¹, Tomofumi Yuki² and Olivier Sentieys²  
¹Université de Rennes / Inria / IRISA, FR; ²INRIA, FR  
**Abstract**  
In this paper, we propose a method to improve the scalability of Word-Length Optimization (WLO) for large applications that use complex quality metrics such as Structural Similarity (SSIM). The input application is decomposed into smaller kernels to avoid uncontrolled explosion of the exploration time, which is known as noise budgeting. The main challenge addressed in this paper is how to allocate noise budgets to each kernel. This requires capturing the interactions across kernels. The main idea is to characterize the impact of approximating each kernel on accuracy/cost through simulation and regression. Our approach improves the scalability while finding better solutions for Image Signal Processor pipeline.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

### 12.4.2 TRADING SENSITIVITY FOR POWER IN AN IEEE 802.15.4 CONFORMANT ADEQUATE DEMODULATOR

**Speaker:** Paul Detterer, Eindhoven University of Technology, NL  
**Authors:** Paul Detterer¹, Cumhur Erdin¹, Jos Huiskens¹, Hailong Jiao¹, Majid Nabi¹, Twan Basten¹ and Jose Pineda de Gyvez²  
¹Eindhoven University of Technology, NL; ²NXP Semiconductors, US  
**Abstract**  
In this work, a design of an IEEE 802.15.4 conformant O-QPSK demodulator is proposed, that is capable of trading off receiver sensitivity for power savings. Such design can be used to meet rigid energy and power constraints for many applications in the Internet-of-Things (IoT) context. In a Body Area Network (BAN), for example, the circuits need to operate with extremely limited energy sources, while still meeting the network performance requirements. This challenge can be addressed by the paradigm of adequate computing, which trades off excessive quality of service for power or energy using approximation techniques. Three different, adjustable approximation techniques are integrated into the demodulation to trade off effective signal quantization bit-width, filtering performance, and sampling frequency for power. Such approximations impact incoming signal sensitivity of the demodulator. For detailed tradeoff analysis, the proposed design is implemented in a commercial 40-nm CMOS technology to estimate power and in Python environment to estimate sensitivity. Simulation results show up to 64% power savings by sacrificing 7 dB sensitivity.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

### 12.4.3 APPROXIMATION TRADE offs IN AN IMAGE-BASED CONTROL SYSTEM

**Speaker:** Sayandip De, Eindhoven University of Technology, NL  
**Authors:** Sayandip De, Sajid Mohamed, Konstantinos Bimpisidis, Dip Goswami, Twan Basten and Henk Corporaal, Eindhoven University of Technology, NL  
**Abstract**  
Image-based control (IBC) systems use camera sensor(s) to perceive the environment. The inherent compute-heavy nature of image processing causes long processing delay that negatively influences the performance of the IBC systems. Our idea is to reduce the long delay using coarse-grained approximation of the image signal processing pipeline without affecting the functionality and performance of the IBC system. The question is: how is the degree of approximation related to the closed-loop quality-of-control (QoC), memory utilization and energy consumption? We present a software-in-the-loop (SiL) evaluation framework for the above approximation-in-the-loop system. We identify the error resilient stages and the corresponding coarse-grained approximation settings for the IBC system. We perform trade off analysis between the QoC, memory utilisation and energy consumption for varying degrees of coarse-grained approximation. We demonstrate the effectiveness of our approach using a concrete case study of a lane keeping assist system (LKAS). We obtain energy and memory reduction of up to 84% and 29% respectively, for 28% QoC improvements.

Download Paper (PDF; Only available from the DATE venue WiFi)

---

### 12.5 Cyber-Physical Systems for Manufacturing and Transportation

**Date:** Thursday 12 March 2020  
**Time:** 16:00 - 17:30  
**Location / Room:** Bayard  
**Chair:** Ulrike Thomas, Chemnitz University of Technology, DE  
**Co-Chair:** Robert De Simone, INRIA, FR

Modeling and design of transportation and manufacturing systems from a cyber-physical system (CPS) perspective have lately attracted extensive attention and the session covers various aspects, from modelling of traffic intersections and control of traffic signals, to implementations of iterative learning controllers for control blocks. Other contributions deal with the selection of network architectures for manufacturing plants and the Digital Twin of production processes for validation.
PARALLEL IMPLEMENTATION OF ITERATIVE LEARNING CONTROLLERS ON MULTI-CORE PLATFORMS

Speaker: Mojtaba Haghi, Eindhoven University of Technology, NL
Authors: Mojtaba Haghi, Yusheng Yao, Dip Goswami and Kees Goossens
1Eindhoven University of Technology, NL; 2Eindhoven university of technology, NL

Abstract
This paper presents design and implementation techniques for iterative learning controllers (ILCs) targeting predictable multi-core embedded platforms. Implementation on embedded platforms results in a number of timing artifacts. Sensor-to-actuator delay (referred to as delay) is an important timing artifact which influences the control performance by changing the dynamic behavior of the system. We propose a delay-based design for ILCs that identifies and operates in the performance-optimal delay region. We then propose two implementation methods -- sequential and parallel -- for ILCs targeting the predictable multi-core platforms. The proposed methods enable the designer to carefully adjust the scheduling to achieve the optimal delay region in the resulting control system. We validate our results by the hardware-in-the-loop (HIL) simulation, considering a motion system as a case-study.

Download Paper (PDF; Only available from the DATE venue WiFi)
12.6.1 WAFER-LEVEL TEST PATTER RECOGNITION AND TEST CHARACTERISTICS FOR TEST-INDUCED DEFECT DIAGNOSIS
Authors: Andrew Yi, Nan Huang, Katherine Shu-Min Li, Ken Chau-Cheng Cheng, Ji-Wei Li, Leon Li, Yang Chen, Nova Cheng, Yen Tsai, Sinya-Wang, Chen-Shun Lee, Leon Chou, Peter Yi-Yu Liao, Hsin-Chung Liang, and Jesu E Chen
NXP Semiconductors Taiwan Ltd., TW; National Sun Yat-sen University, TW; NXP Semiconductors Taiwan Ltd., TW; National Sun Yat-Sen University, TW; National Chung-Hsing University, TW; 4Chung Yuan Christian University, TW; National Central University, TW
Abstract
Wafer defect maps provide precious information of fabrication and test process defects, so they can be used as valuable sources to improve fabrication and test yield. This paper applies artificial intelligence based pattern recognition techniques to distinguish fab-induced defects from test-induced ones. As a result, test quality, reliability and yield could be improved accordingly. Wafer test data contain site-dependent information regarding test configurations in automatic test equipment, including effective load push force, gap between probe and loadboard, probe tip size, probe-cleaning force, etc. Our method analyzes both the test paths and site-dependent test characteristics to identify test-induced defects. Experimental results achieve 96.83% prediction accuracy of six different NXP products, which show that our methods are both effective and efficient.
Download Paper (PDF; Only available from the DATE venue WiFi)

12.6.2 A METHOD OF VIA VARIATION INDUCED DELAY COMPUTATION
Authors: Moonsu Kim, Yun Heo, Seungjae Jung, Kelvin Lee, Youngmin Shin, Hathaniel Conos, and Hafat Fatemi
Samsung, KR; Synopsys, US
Abstract
As process technologies are scaled down, interconnect delay becomes a major component of entire path delay, and vias represent a significant portion of the interconnect delay. In this paper, we present a novel variation-aware delay computation method for vias is proposed. Our experiments show that this method can reduce over five percent of pessimism in arrival time calculation when it is compared with state-of-the-art solutions.
Download Paper (PDF; Only available from the DATE venue WiFi)

12.6.3 FULLY AUTOMATED ANALOG SUB-CIRCUIT CLUSTERING WITH GRAPH CONVOLUTIONAL NEURAL NETWORKS
Speaker: Keertana Settlaui, University of California, Berkeley, US
Authors: Keertana Settlaui and Elias Fallon
University of California, Berkeley, US; Cadence Design Systems, US
Abstract
The design of custom analog integrated circuits is one of the contributing factors in high development cost and increased production time, driving the need for more automation in this space. In automating particular avenues of analog design, it is then crucial to assess the efficacy with which the automation algorithm is able to solve the desired problem. To do this, one must consider four metrics that are especially pertinent in this area: robustness, accuracy, level of automation, and computation time. In this work, we present a framework that bridges the gap between schematic and layout generation, by encapsulating the design intuition needed to create layout through identification of critical sub-circuit structures. Our approach focuses on identifying analog sub-circuits by utilizing Graphical Convolutional Neural Networks (GCNNs) in conjunction with an unsupervised graph clustering technique to result in the first tool, to our knowledge, to entirely automate this clustering process. We compare our algorithm to prior work in this space utilizing the four important figures of merit, and our results show over 90% accuracy across six different analog circuits, ranging in size and complexity, while taking just under 1 second to complete.
Download Paper (PDF; Only available from the DATE venue WiFi)

12.6.4 EVPS: AN AUTOMOTIVE VIDEO ACQUISITION AND PROCESSING PLATFORM
Speaker: Christophe Flouza, CEA LIST, FR
Authors: Christophe Flouza, Erwan Priou, Mickael Guibert, Bojan Jovanovic, and Mohamad Oussayyane
CEA LIST, FR; CEa List, FR
Abstract
This paper describes a versatile and flexible video acquisition and processing platform for automotive. It is designed to meet aggressive requirements in terms of bandwidth and latency when implementing ADAS functions. Based on the Xilinx Ultrascale+ FPGA device, a vision processing pipeline mixing software and hardware tasks is implemented on this platform. This setup is able to collect four automotive camera streams (MIPI CSII) and process them in the loop before transmitting a more intelligible pre-processed/enhanced data.
Download Paper (PDF; Only available from the DATE venue WiFi)

12.6.5 AN ON-BOARD ALGORITHM IMPLEMENTATION ON AN EMBEDDED GPU: A SPACE CASE STUDY
Speaker: Ivan Rodriguez, UPC / BSC, ES
Authors: Ivan Rodriguez, Leonidas Kosmidis, Olivier Notebaert, Francisco J Cazorla, and David Steenari
UPC / BSC, ES; IBSAC, ES; Airbus Defence and Space, FR; European Space Agency, NL
Abstract
On-board processing requirements of future space missions are constantly increasing, calling for new hardware than the traditional ones used in space. Embedded GPUs are an attractive candidate offering both high performance capabilities and low power consumption, but there are no complex industrial case studies from the space domain demonstrating these advantages. In this paper we present the GPU parallelisation of an on-board algorithm, as well as its performance on a promising embedded GPU COTS platform targeting critical systems.
Download Paper (PDF; Only available from the DATE venue WiFi)

12.6.6 TLS-LEVEL SECURITY FOR LOW POWER INDUSTRIAL IOT NETWORK INFRASTRUCTURES
Authors: Jochen Mades, Gerd Ebert, Boris Janjic, Frederik Lauer, Carl Rheinländer, and Norbert Wehn
KSB SE Co. KGaA, DE; TU Kaiserslautern, DE
Abstract
The Industrial Internet of Things (IIoT) enables communication services between machinery and cloud to enhance industrial processes e.g. by collecting relevant process parameters or providing predictive maintenance. Since the data is often origin from critical infrastructures, the security of the data channel is the main challenge, and is often weakened due to limited compute power and energy availability of battery-powered sensor nodes. Lightweight alternatives to standard security protocols avoid computationally intensive algorithms, however, they do not provide the same level of trust as established standards such as Transport Layer Security (TLS). In this paper, we propose an IIoT network system that enables a secure end-to-end IP communication between ultra-low-power sensor nodes and cloud servers. It provides full TLS support to ensure perfect forward secrecy by using hardware accelerators to reduce the energy demand of the security algorithms. Our results show that the energy overhead of the TLS handshake can be significantly reduced to enable a secure IIoT infrastructure with a reasonable battery lifetime of the edge devices.
Download Paper (PDF; Only available from the DATE venue WiFi)

17:30 End of session
12.7 Power-efficient multi-core embedded architectures

Date: Thursday 12 March 2020
Time: 16:00 - 17:30
Location / Room: Berlioz

Chair: Andreas Burg, EPFL, CH
Co-Chair: Semeen Rehman, TU Wien, AT

This session has papers that provide power-efficiency solutions for multi-core embedded architectures. Techniques discussed in the session are related to the architectural measures as well as effectively controlling voltage-frequency settings using machine learning based on user experiences.

<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:00</td>
<td>12.7.1</td>
<td>TUNING THE ISA FOR INCREASED HETEROGENEOUS COMPUTATION IN MPSoCs</td>
</tr>
<tr>
<td>Authors: Pedro Henrique Exenberger Becker, Jackson Dellagostin Souza and Antonio Carlos Schneider Beck, Universidade Federal do Rio Grande do Sul, BR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abstract: Heterogeneous MPSoCs are crucial to meeting energy efficiency and performance, given their combination of cores and accelerators. In this work, we propose a novel technique for MPSoCs design, increasing their specialization and task-parallelism within a given area and power budget. By removing the microarchitectural support of costly ISA extensions (e.g., FP, SIMD, crypto) from a few cores (transforming them into Partial-ISA Cores), we make room to add extra (full and simpler) in-order cores and hardware accelerators. While applications must migrate from Partial-ISA cores when they need the removed ISA support, they also execute at lower power consumption during their ISA-extension-free phases, since partial cores have much simpler datapaths compared to their full-ISA counterparts. On top of it, the additional cores and accelerators increase task-level parallelism and make the MPSoC more suitable for application-specific scenarios. We show the effectiveness of our approach by composing different MPSoCs in distinct execution scenarios, using the FP instructions and RISC-V ISA as a case study. To support our system, we also propose two scheduling policies, performance- and energy-oriented, to coordinate the execution of this novel design. For the former policy, we achieve 2.8x speedup for a neural network road sign detection, 1.53x speedup for a video-streaming app, and 1.2x speedup for a task-parallel scenario, consuming 68%, 75%, and 33% less energy, respectively. For the energy-oriented policy, partial-ISA reduces energy consumption by 29% over a highly efficient baseline, with increased performance.

Download Paper (PDF; Only available from the DATE venue WiFi)

| 16:30 | 12.7.2 | USER INTERACTION AWARE REINFORCEMENT LEARNING FOR POWER AND THERMAL EFFICIENCY OF CPU-GPU MOBILE MPSoCs |
| Speaker: Somdip Dey, University of Essex, GB |
| Authors: Somdip Dey, Amit Kumar Singh, Xiaohang Wang and Klaus McDonald-Maier |
| University of Essex, GB; South China University of Technology, CN |

Abstract: Mobile user’s usage behaviour changes throughout the day and the desirable Quality of Service (QoS) could thus change for each session. In this paper, we propose a QoS aware agent to monitor mobile user’s usage behaviour to find the target frame rate, which satisfies the desired user’s QoS, and applies reinforcement learning based DVFS on a CPU-GPU MPSoC to satisfy the frame rate requirement. Experimental study on a real Exynos hardware platform shows that our proposed agent is able to achieve a maximum of 50% power saving and 29% reduction in peak temperature compared to stock Android’s power saving scheme. It also outperforms the existing state-of-the-art power and thermal management scheme by 41% and 19%, respectively.

Download Paper (PDF; Only available from the DATE venue WiFi)

| 17:00 | 12.7.3 | ENERGY-EFFICIENT TWO-LEVEL INSTRUCTION CACHE DESIGN FOR AN ULTRA-LOW-POWER MULTI-CORE CLUSTER |
| Speaker: Jie Chen, Università di Bologna, CN |
| Authors: Jie Chen, Igor Loi, Luca Benini and Davide Rossi |
| University of Bologna, GB; GreenWaves Technologies, FR; Università di Bologna, IT |

Abstract: High Energy efficiency and high performance are the key regiments for Internet of Things (IoT) edge devices. Exploiting clusters of multiple programmable processors has recently emerged as a suitable solution to address this challenge. However, one of the main power bottlenecks for multi-core architectures is the instruction cache memory. We propose a two-level structure based on Standard Cell Memories (SCMs) which combines a private instruction cache (L1) per-core and a low-latency (only one cycle latency) shared instruction cache (L1,S). We present a detailed comparison of performance and energy efficiency for different instruction cache architectures. Our system-level analysis shows that the proposed design improves upon both state-of-the-art private and shared instruction cache architectures and balances well performance with energy-efficacy. On average, when executing a set of real-life IoT applications, our multi-level cache improves performance and energy efficiency both by 10% with respect to the private instruction cache system, and improves energy efficiency by 15% and 7% with a performance loss of only 2% with respect to the shared instruction cache. Besides, relaxed timing makes two-level instruction cache an attractive choice for aggressive implementation, with more slack for convergence in physical design.

Download Paper (PDF; Only available from the DATE venue WiFi)

17:30 End of session

12.8 Special Session: EDA Challenges in Monolithic 3D Integration: From Circuits to Systems

Date: Thursday 12 March 2020
Time: 16:00 - 17:30
Location / Room: Exhibition Theatre

Chair: Pascal Vivet, CEA-Leti, FR
Co-Chair: Mehdi Tahoori, Karlsruhe Institute of Technology, DE

Monolithic-3D integration (M3D) has the potential to improve the performance and energy efficiency of 3D ICs over conventional TSV-based counterparts. By using significantly smaller inter-layer vias (ILVs), M3D offers the ‘true’ benefits of utilizing the vertical dimension for system integration: M3D provides ILVs that are 100x smaller than a TSV and have similar dimensions as normal vias in planar technology. This allows M3D to enable high-performance and energy-efficient systems through higher integration density, flexible partitioning of logic blocks across multiple layers, and significantly lower total wire-length. From a system design perspective, M3D is a breakthrough technology to achieve “More Moore and More Than Moore,” and opens up the possibility of creating manycore chips with multi-tier cores and network routers by utilizing ILVs. Importantly, this allows us to create scalable manycore systems that can address the communication and computation needs of big data, graph analytics, and other data-intensive parallel applications. In addition, the dramatic reduction in via size and the resulting increase in density opens up numerous opportunities for design optimizations in the manycore domain.
<table>
<thead>
<tr>
<th>Time</th>
<th>Label</th>
<th>Presentation Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:00</td>
<td>12.8.1</td>
<td>M3D-ADTCO: MONOLITHIC 3D ARCHITECTURE, DESIGN AND TECHNOLOGY CO-OPTIMIZATION FOR HIGH ENERGY-EFFICIENT 3D IC</td>
<td>Sebastien Thuries, Olivier Billoint, Sylvain Choisent, Didier Lattard, Romain Lemaire and Perrine Batude, CEA-Leti, FR</td>
</tr>
<tr>
<td>16:30</td>
<td>12.8.2</td>
<td>DESIGN OF A RELIABLE POWER DELIVERY NETWORK FOR MONOLITHIC 3D ICS</td>
<td>Krishnendu Chakrabarty, Duke University, US</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker: KL Krishna Chakrabarty, Duke University, US</td>
<td>Authors: Shao-Chun Hung and Krishnendu Chakrabarty, Duke University, US</td>
</tr>
<tr>
<td>17:00</td>
<td>12.8.3</td>
<td>POWER-PERFORMANCE-THERMAL TRADE-OFFS IN M3D-ENABLED MANYCORE CHIPS</td>
<td>Partha Pande, Washington State University, US</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker: Partha Pande, Washington State University, US</td>
<td>Authors: Shouvik Musavvir¹, Anwesha Chatterjee ¹, Ryan Kim², Daehyun Kim¹, Janardhan Rao Doppa¹ and Partha Pratim Pande ¹</td>
</tr>
<tr>
<td></td>
<td></td>
<td>¹Washington State University, US; ²Colorado State University, US</td>
<td><strong>Abstract</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Monolithic 3D (M3D) technology enables unprecedented degrees of integration on a single chip. The miniscule monolithic inter-tier vias (MIVs) in M3D are the key behind higher transistor density and more flexibility in designing circuits compared to conventional through silicon via (TSV)-based architectures. This results in significant performance and energy-efficiency improvements in M3D-based systems. Moreover, the thin inter-layer dielectric (ILD) used in M3D provides better thermal conductivity compared to TSV-based solutions and eliminates the possibility of thermal hotspots. However, the fabrication of M3D circuits still suffers from several non-ideal effects. The thin ILD layer may cause electrostatic coupling between tiers. Furthermore, the low-temperature annealing degrades the top-tier transistors and bottom-tier interconnects. An NoC-based manycore design needs to consider all these M3D-process related non-idealities. In this paper, we discuss various design challenges for an M3D-enabled manycore chip. We present the power-performance-thermal trade-offs associated with these emerging manycore architectures.</td>
<td>Download Paper (PDF; Only available from the DATE venue WiFi)</td>
</tr>
<tr>
<td>17:30</td>
<td></td>
<td>End of session</td>
<td><strong>Source URL:</strong> <a href="https://past.date-conference.com/date20/booklet/proof_reading">https://past.date-conference.com/date20/booklet/proof_reading</a></td>
</tr>
</tbody>
</table>