IEEE VLSI Circuits and Systems Letter

Volume 8, Issue 3, Aug 2022

Editorial

Invited articles

- “Making Hardware Design Look More Like Software Design – What Are The Missing Pieces?” by Dr. Rob Aitken, Director of Office of Technology and Strategy, Synopsys
- “Scalable Algorithmic Checks for Autonomous Systems: Electronics, Control and Machine Intelligence” by Prof. A. Chatterjee and C. Amarnath, Georgia Institute of Technology

Conference spotlight

- ISVLSI 2022 – Report by Conference General Co-Chairs – Theocharis Theocharides (Univ of Cyprus), Muhammad Shafique (NYU Abu Dhabi), Hai "Helen" Li (Duke University)
- ISVLSI 2022 Best Paper “Efficient Accelerator of Deformable 3D Convolutional Network for Video Super-Resolution” - Siyu Zhang, Wendong Mao and Zhongfeng Wang (Nanjing University, China)
- ISVLSI-2022 Best Paper “Adaptable Multi-Level Voltage to Binary Converter Using Ferroelectric FETs” - Sanjay Das and Arun Govindankutty (North Dakota State University, USA), by Shan Deng (Rochester Institute of Technology, USA), by Kai Ni (RIT, USA) and Sumitha George (North Dakota State University, USA)

Women in VLSI (WiV) series spotlight

- Interview with Dr. Diana Marculescu, Department Chair of the Electrical and Computer Engineering Department at UT Austin, USA.

Updates

- Recent Announcements of Interest – Ishan Thakkar
- Planned 2022 conference sponsorships by TCVLSI
From the Editor-in-Chief’s Desk - Editorial

The IEEE VLSI Circuits and Systems Letter (VCaSL) is affiliated with the Technical Committee on VLSI (TCVLSI) under the IEEE Computer Society. It aims to report recent advances in VLSI technology, education, and opportunities and, consequently, grow the research and education activities in the area. The letter published quarterly (since 2018), highlights snippets from the vast field of VLSI including semiconductor design, digital circuits and systems, analog and radio-frequency circuits, as well as mixed-signal circuits and systems, logic, microarchitecture, architecture and applications of VLSI. TCVLSI aims to encourage efforts around advancing the field of VLSI be it in the device, logic, circuits or systems space, promoting secured computer-aided design, fabrication, application, and business aspects of VLSI while encompassing both hardware and software.

IEEETCVLSI sponsors a number of premium conferences and workshops, including, but not limited to, ASAP, ASYNC, ISVLSI, IWLS, SLIP, and ARITH. Emerging research topics and state-of-the-art advances on VLSI circuits and systems are reported at these events on a regular basis. Best paper awards are selected at these conferences to promote the high-quality research work each year. In addition to these research activities, TCVLSI also supports a variety of educational activities related to TCVLSI. Typically, several student travel grants are sponsored by TCVLSI at the following conferences: ASAP, ISVLSI, IWLS, iSES (formerly iNIS) and SLIP. Funds are typically provided to compensate student travels to these conferences as well as to attract more student participation. The organizing committees of these conferences undertake the task of selecting right candidates for these awards.

This issue of VCaSL features two invited articles. The first “Making Hardware Design Look More Like Software Design – What Are The Missing Pieces?” by Dr. Rob Aitken from Synopsys, discusses why hardware design today does not look like software design and what will it take to get us there. The second “Scalable Algorithmic Checks for Autonomous Systems: Electronics, Control and Machine Intelligence” by Prof Abhijit Chatterjee and Chandramouli Amarnath of Georgia Tech calls for hierarchical error-checking methodologies running in real-time to minimize and eliminate operational hazards in systems like autonomous robots and self-driving cars.

The newsletter spotlights one of TCVLSI’s sponsored conferences in 2022, the IEEE Computer Society Annual Symposium on VLSI (ISVLSI). One-page teasers of two best papers awarded at the 2022 ISVLSI conference are showcased: “An Efficient Accelerator of Deformable 3D Convolutional Network for Video Super-Resolution” and “Adaptable Multi-Level Voltage to Binary Converter Using Ferroelectric FETs”.

In our Women in VLSI (WiV) series, we share an inspiring interview with Prof. Diana Marculescu, Department Chair of the Electrical and Computer Engineering Department at UT Austin, USA.

Additionally, included is a section on relevant recent announcements collated by our Associate Editor, Ishan Thakkar.

I’d like to thank Dr. Olivier Franza for designing the cover page of this newsletter. Thank you to the authors of the various articles. I’d like to thank the IEEE CS staff, for their professional services to make the newsletter publicly available. I’d love to hear from the readers on what you would like to see in future newsletters. I welcome recommendations/feedback via email. Happy reading.

Mondira (Mandy) Pant, Ph.D
Chair, IEEE Computer Society TCVLSI
Editor-in-Chief of IEEE VCaSL, TCVLSI
Intel Corporation, USA

IEEE CS-TCVLSI: https://www.computer.org/communities/technical-committees/tcvlsi
Email: mondira.pant@ieee.org

TCVLSI has a total of about 1000 active members as of Aug 2021 and a readership of about 30,000
To join TCVLSI (its free), click here: https://www.ieee.org/membership-catalog/productdetail/showProductDetailPage.html?product=CMYVLSI732

https://tc.computer.org/tcvlsi/
Making Hardware Design Look More Like Software Design
– What Are The Missing Pieces?

Rob Aitken. Synopsys Inc < rob.aitken@synopsys.com>

The idea of making hardware design look more like software design is not new. The principles of silicon compilation, as articulated by Dave Johannsen in his DAC 1979 paper [1], show thinking along those lines more than 40 years ago, including Johannsen’s objectives of building from a library of standard components, abstracting away as much physical detail as possible, and allowing for experimentation across a wide space of potential implementations before settling on a solution. Similar thoughts underlie efforts in high level synthesis – generating hardware from a behavioral description of its function, typically written in a high-level programming language such as C++ – which already had a substantial body of literature behind it by the late 1980s [2].

In an idealized view, software design begins with identification of requirements, followed by an exploration of algorithms and architectures, which are in turn implemented, tested, deployed and maintained. Well defined abstraction layers, extensive use of libraries, strongly-typed languages and so on ensure a smooth process. Hardware design, especially chip design, by contrast, struggles to move to abstraction levels higher than logic gates and memories. Immense amounts of effort are devoted to handling details we wish were abstracted away, such as power supply integrity and clock management. What if hardware design could be more like software design?

Of course, there are good reasons why hardware design does not mimic software design, beginning with economics. A software executable can be distributed worldwide for virtually no cost, and all the copies are identical. Chips, on the other hand, are physical objects. Every copy costs money to create and takes weeks or months to be manufactured. The copies themselves are not identical – many manufactured chips do not work, and sometimes their failure modes are very subtle and difficult to detect. A software bug can be fixed by distributing a new code release. Hardware bugs can sometimes be fixed with a software workaround, but many require a respin of the design, with a significant cost in money and time.

Still, the vision of a better approach to hardware design remains appealing. High-level synthesis tools are routinely used for certain classes of designs, especially those implementing algorithmic transformations like encryption/decryption or video decoding. The Agile Hardware work at Stanford [3] and RISC-V/Chisel ecosystem originally developed at Berkeley [4] are based on making the boundary between hardware and software less abrupt. These approaches are primarily targeted at accelerators and CPUs respectively, but each is a step towards the larger goal.

A full discussion of the topic is well beyond the scope of this article, but decades of work have provided us with both observations and promising directions. At a minimum, the hardware implementations that come from a software-defined flow need to be economically viable – if they have an area, power, or performance cost beyond that of traditional implementation, that cost needs to be manageable. In addition, the physical aspects of implementation (power delivery, timing signoff, etc.) along with standard add-ons such as DFT need to “just work” without significant intervention. Minor changes in input cannot lead to large, unpredictable changes in output. Beyond that, the software that will be converted into hardware needs to have an inherently parallel structure that is able to mimic the independence and concurrence of operations in physical chip hierarchies. It also needs to be amenable to high level architecture choices – CPU versus GPU, two cores versus eight, shared memory or separate – as well as classic hardware tradeoffs including...
speed versus flexibility and optimization for performance, power or some combination of the two. In other words, the architects and writers of the hardware-generating software need to be able to convey their intentions to the implementation tools, the tools need to respect these intentions and be able to follow them, and designers need to be able to iterate with the flow and tune the results in predictable ways.

Design intent needs to travel through the entire tool flow, starting with logic synthesis, so that future optimizations do not negate past ones, and so that subsequent refinements of the initial design will lead to the desired behavior of the resulting hardware, rather than landing in some arbitrary new point in design space. Existing systems struggle with preserving design intent. A user of a parameterized RTL generator, for example, may have chosen 4 parallel operations instead of 8, but if a record of that choice is nowhere to be found in the generated RTL, then a subsequent synthesis tool may conclude that it could improve performance by doubling the hardware for each of the 4 operations, quietly negating the designer’s desired choice. Preserving designer intent is also important for both verification (“does the implementation do what it was meant to do?”) and security analysis (“does the implementation do anything it is not meant to do?”).

Finding the next abstraction layer beyond RTL remains an open problem, but one promising candidate is the CIRCT (Circuit IR Compilers and Tools) project [5] based on MLIR (Multi-Level Intermediate Representation) [6]. MLIR is primarily intended for software compilation, and provides several mechanisms to preserve design intent, including a required connection between a representation and a line in the original source code where it originated, and the concept of “progressive lowering” which allows transformations to lower levels to occur for individual elements in incremental steps while multiple abstraction levels coexist simultaneously. The CIRCT project is focused on hardware design and contains several dialects including representations for FIRRTL [4] and SystemVerilog.

Another important challenge for today’s hardware and software lies in scale. Managing design hierarchy is a key aspect of successful chip implementation. While some classes of tool (e.g. timing signoff) can operate at the chip level, with tens of billions of transistors, many can only operate at a block or subsystem level. Existing software-style approaches to hardware design work for accelerator sized blocks but not for full chips. Combining chip-scale hardware design and cloud-scale software development is an ongoing challenge that will require new approaches to be successful. Gannon et al [7] describe some key properties of truly cloud native software: global scale in both reach and number of users, inherent fault tolerance, continuous uptime including test and upgrades, and fully integrated security. Cloud-native applications themselves are containerized, dynamically orchestrated and microservice oriented. Most existing EDA algorithms were developed with an implicit compute model of a monolithic application, running on one or a few processors, using shared memory space and/or an NFS file system to move information between processes. Architecting tools that innately take advantage of cloud computing’s capabilities can potentially enable solutions at a scale that is unachievable now.

It may be that hardware design is too different from software design for the former to ever become a simple offshoot or variation of the latter, but hardware design continues to borrow and adapt techniques from software design. This work is valuable and worth continuing, not just to make hardware design better and easier, but also to pave the way for broader visions, including true hardware/software co-design.
Rob Aitken leads the Office of Technology Strategy at Synopsys, where he and his team are looking into future technical opportunities for the company, including future design methodologies. Prior to joining Synopsys in 2022 he worked at Arm, Artisan, HP and Agilent. He is an inventor on over 50 US patents and an author of over 100 publications on a variety of topics. Dr. Aitken is an IEEE Fellow and has served as General Chair of DAC and ITC.
Scalable Algorithmic Checks for Autonomous Systems: Electronics, Control and Machine Intelligence

A. Chatterjee and C. Amarnath, School of ECE, Georgia Tech

Safety, security and trustworthiness have become common buzzwords for emerging autonomous systems due to their proliferation in society, widespread use and safety-criticality. Autonomous robots and self-driving cars, for example, demand high levels of resilience to computation errors in signal processing and control subsystems as well as performance degradation and failures in embedded sensors, actuators and associated electronics. In addition, rapid detection and mitigation of external security attacks either through injection of malware or hardware trojans is of paramount importance for high system trustworthiness. The inherent complexity of emerging autonomous systems calls for a hierarchical error-checking methodology that can scale across real-time electrical subsystems such as digital and analog filters, electro-mechanical sensors, electrical motors and other actuators and all relevant support electronics and control systems, both digital and analog. Most important, intelligent systems employ deep learning algorithms. These also need to be checked in real-time to minimize and eliminate operational hazards.

The origins of this research trace back to work on algorithmic fault tolerance which focused on concurrent error detection and mitigation in linear signal processing algorithms such as for matrix multiplication, fast fourier transform and digital filtering [1-2]. Of specific interest is encoded matrix-vector multiplication $[Y]=[A][X]$, where $[A]$ is an $n$ by $m$ matrix of real numbers and $[Y]$ and $[X]$ column are vectors with dimensions $n$ and $m$, respectively. To encode the matrix-vector multiplication, an additional row $[R]$ is added to $[A]$ and an additional element $[C]$ is added to $[Y]$ to give $\begin{bmatrix} Y \\ C \end{bmatrix} = \begin{bmatrix} A \\ R \end{bmatrix} \begin{bmatrix} X \end{bmatrix}$ (Equation (1)), where $R$ is an $m$-dimensional row vector given by $[CV],[A]$ and $[CV] = [\alpha_1, \alpha_2, \ldots, \alpha_n]$ is an $n$-dimensional vector called the coding vector. It can be shown that under such encoding, $[C] = [CV].[Y]$ ([C] is called the check variable). Hence, the difference $[CV].[Y]-[C]$ can be used as a check and is ideally zero in the absence of errors in matrix-vector computations and non-zero otherwise.

A key realization in the application of the above matrix encoding techniques to error detection in physical electronic systems is that physical systems such as filters, motors and robots execute algorithms dictated by principles of physics. For example, the speed of a DC motor is a function of the applied voltage, flux and armature resistance of the motor. The equations that dictate the motor speed implicitly define the physics-based algorithm executed by the DC motor. Similarly, the laws of physics determine the control principles necessary for balancing an inverted pendulum attached to a moving cart, essentially defining an algorithm for the balancing process. If it is possible to find matrix representations of such algorithms, similar to that described for matrix-vector multiplication above, then real-time error detection in physical systems is possible by encoding those representations in appropriate ways.

Such representations called state variable representations have been well studied in the literature and permit a pathway to applying algorithmic checks to large classes of everyday electronic system. State variable representations are typically given by discrete time descriptions of the form $X(t+1)=[A].X(t)+[B].U(t)$, where $X(t)$ is the system state vector at time $t$ and $U(t)$ is the system input vector at time $t$. In analog form, $X'(t)=[A].X(t)+[B].U(t)$, where $X(t)$ is the time derivative of the state vector $X(t)$. By encoding the state and input matrices $[A]$ and $[B]$ respectively [2,3] , as described earlier for matrix-vector multiplication, concurrent error detection can be performed for real-time systems using algorithmic state-space checks. These errors can consist of transient errors in electronic computation or changes in the entries of the $[A]$ and $[B]$ matrices induced by performance degradation, wearout and failure of the
underlying electronics and electro-mechanical sensors and actuators. In addition, the checks can be used to detect malicious external security attacks and hardware trojans with carefully engineered protection for the checks themselves. In the following, we give a few examples of real-time systems to which such checking techniques can be applied. A key advance is the extension of linear checksums above to nonlinear checks which allow concurrent error detection in broad classes of nonlinear systems. For such systems, the computation of the check variable \([C]\) is determined as a nonlinear transform of \([X]\) in Equation (1). This nonlinear transform is learned using machine-learning algorithms from experiments conducted on the nonlinear real-time system.

**State Space Check - Analog Circuit:** Concurrent error detection for a 6th order Butterworth analog filter is shown in Figure 1 [4]. The filter consists of 6 opamps and the error checking circuit is constructed out of 3 opamps. Figure 1 (a) shows a sinusoidal filter input waveform and the corresponding filter output. In Figure 1 (b), a signal perturbation (noise) is used to corrupt one of the filter state variables producing a corrupted output. For the same perturbation, Figure 1 (c) shows how the non-zero error signal detects the injected error almost immediately after error onset. It is possible to diagnose and correct the error using phase-inverted error feedback and one additional check. Figure 1 (d) shows the corrected filter output for the same state perturbation.

**State Space Check - DC Motor Used for Servo-Control:** Figure 2 shows how state space checks can be applied to a DC motor used in a servomotor configuration [5]. Figure 2 (b) shows the state variable representation for the DC servo-motor. This representation is encoded and additional electronics with an opamp is used to implement the checking circuitry. Figure 2 (c) shows how the error signal, which is ideally zero for the fault-free system, becomes instantaneously non-zero when a power supply transient is
experienced by the motor. Any change in the torque-current relationship of the motor is detected instantaneously by the check,

**Algorithmic Check - Automotive Brake-by-Wire System:** A key advance is the application of algorithmic checks to nonlinear state variable systems. The state space representation for a brake-by-wire systems is complex but allows computation of the checksum of the system’s state at the current time from a nonlinear transform of prior state values and system inputs over a finite time window [6]. Figure 3 (b) shows wheel lock-up due to a soft error injected into the brake-by-wire controller software. Figure 3 (c) shows how this is detected instantaneously by the resulting check for the system.

**Algorithmic Check - Deep Learning Network:** Algorithmic checks can further be applied to deep learning systems such as convolutional neural networks (CNNs) [7], as shown in Figure 4(a). Linear transforms such as the Discrete Cosine Transform are used to encode the inputs to the layer computations (Step 1 and 2) as well as layer outputs (Steps 4 and 5) to predict a linear function of layer outputs (encoded layer outputs) (Step 3), detecting errors in computation based on statistical distributions of prediction error (Step 6) over training data (Step 7). Figure 4(b) shows this applied to activation error detection in dense layers of a learning systems with high coverage. The framework applies to convolutional layer computations as well and can be adapted to detect security attacks.
References:


Abhijit Chatterjee is a Professor in the School of Electrical and Computer Engineering at Georgia Tech and a Fellow of the IEEE. He received his Ph.D in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 1990. Dr. Chatterjee received the NSF Research Initiation Award in 1993 and the NSF CAREER Award in 1995. He has received seven Best Paper Awards and three Best Paper Award nominations. His work on self-healing chips was featured as one of General Electric’s key technical achievements in 1992 and was cited by the *Wall Street Journal*. In 1995, he was named a Collaborating Partner in NASA’s New Millennium project. In 1996, he received the Outstanding Faculty for Research Award from the Georgia Tech Packaging Research Center, and in 2000, he received the Outstanding Faculty for Technology Transfer Award, also given by the Packaging Research Center. In 2007, his group received the Margarida Jacome Award for work on VIZOR: Virtually Zero Margin Adaptive RF from the Berkeley Gigascale Research Center (GSRC). Dr. Chatterjee has authored over 450 papers in refereed journals and meetings and has 22 patents. He is a co-founder of Ardext Technologies Inc., a mixed-signal test solutions company and served as chairman and chief scientist from 2000-2002. His research interests include error-resilient machine learning, signal processing and control systems, mixed-signal/RF/multi-GHz design and test and adaptive real-time systems.

Chandramouli Amarnath is a PhD student in the Smart Resilient Systems Lab at Georgia Tech, working with Prof. Abhijit Chatterjee. He received his B.Tech from the National Institute of Technology Karnataka, Surathkal in 2018 and began his PhD at Georgia Tech in 2018. His research interests include fault-tolerance in autonomous systems and deep learning systems as well as security of deep neural networks. His published work addresses topics such as hierarchical failure modeling and recovery in autonomous vehicles, secure and safe operation of deep neural networks and error resilience in state estimation.
ISVLSI 2022

Theocharis Theocharides, Muhammad Shafique, Hai "Helen" Li
General Chairs of ISVLSI 2022

It is our distinct privilege to report about the 2022, and in particular the first in-person edition of the IEEE Computer Society Annual Symposium on VLSI (ISVLSI, http://www.ieee-isvlsi.org/) following the COVID-19 pandemic. The main goal of ISVLSI is to explore emerging trends and novel ideas and concepts in the area of VLSI circuits and systems. ISVLSI provides a platform for both academic and industrial researchers to interact under one roof for research and development which may lead to realization of efficient, robust, intelligent, and secure VLSI circuits and systems. Like its predecessors, ISVLSI 2022 is a sponsored meeting of the Technical Committee on VLSI (TCVLSI, http://www.ieee-tcvlsi.org/), of the IEEE Computer Society (IEEE-CS).

Since early 2020, our community has been facing an unprecedented situation due to the COVID19 pandemic, with the 2020 and 2021 editions of ISVLSI taking place virtually, with participants unfortunately missing on the unparalleled experience of face-to-face interaction that our community strives for. The ISVLSI 2022 organizing committee, taking into consideration the encouraging situation with regards to the pandemic, and the gradual reopening of international travel, has decided to proceed with an in-person event, where attendees will experience once again the fruitful discussions and networking activities, while also given the opportunity to witness the technical contributions of ISVLSI 2022 authors via physical attendance. Obviously, given that we are still facing uncertainty with travel disruptions, as we are just emerging for the effects of the pandemic, we have also decided to offer the opportunity for speakers unable to travel to the venue, to present their work remotely, while engaging with the attendees and answering questions via live remote participation. Given that this maybe a new trend for our community, the organizing committee has also reshuffled the conference format, enhancing it with several interactive features to ensure that all attendees make the most of their time, such as focused plenary talks, a panel, and an embedded tutorial. Social activities have also returned; after all, the social part of the conference is what brings together researchers, academics, industry professionals and business entrepreneurs in our community together to discuss, analyze, evaluate, and identify collaboration opportunities.

Over more than two decades the ISVLSI has been a unique forum exploring emerging trends and novel ideas and concepts in VLSI, while promoting multidisciplinary research and new visionary approaches in VLSI. The Symposium has been successful in bringing together leading international scientists and researchers from academia and industry. ISVLSI is a sister conference of a league of successful meetings such as ARITH, ASAP, iSES, IWLS, and SLIP which are sponsored by the Technical Committee on VLSI (TCVLSI, www.ieee-tcvlsi.org/), of IEEE Computer Society (IEEE-CS).

ISVLSI 2022 covered a wide range of topics: From VLSI circuits, systems and design methods to system-level design and system-on-chip (SoC) issues, to bringing VLSI experience to new areas, architectures, and technologies. Future design methodologies as well as new Electronic Design Automation (EDA) tools to support them are also key topics. ISVLSI 2022 features three separate paper categories: regular papers, poster papers and special session papers.

Regular submissions have been received from across the globe. In total, 93 regular track full submissions
were received (after filtering out the incomplete submissions, i.e., with abstracts only), out of which 37 high-quality papers have been accepted for the oral presentation and the proceedings, thus, making an overall acceptance rate of 39.8%. The regular track papers are divided into 12 oral sessions in the technical program, and each paper is allocated 25 minutes for presentation. All submitted papers underwent a rigorous double-blind-review process by the program committee members and external reviewers, coordinated by the co-chairs per track who represented a strong team of international leading experts in their respective fields. Apart from the 37 technical papers, 11 submissions were selected to be presented as posters. Regular papers received 6 pages in the proceedings, so that the technical depth and quality of the reviewed articles is preserved. Poster papers received 4 pages, as they represent work in progress with great potential.

A number of special tracks are also featured in ISVLSI 2022. In particular, the following 5 tracks offer presentations and papers organized in 12 special sessions of the technical program in cutting-edge topics: Approximate Computing for Machine Learning, Processing-In-Memory, Secured Neuromorphic Computing, Efficient Testing of AI Accelerators, and Secure and Dependable Cyber-Physical Systems.

The program of ISVLSI 2022 also featured 3 excellent keynotes by top experts from industry and academia, including Alberto L. Sangiovanni-Vincentelli (Edgar L. and Harold H. Buttner Chair of Electrical Engineering and Computer Sciences at the University of California at Berkeley), Giacomo Indiveri (Professor at the Faculty of Science of the University of Zurich and at Department of Information Technology and Electrical Engineering of ETH Zurich), and Elisabetta Farella (Head of the Energy Efficient Embedded Digital Architectures (E3DA) unit at the Fondazione Bruno Kessler in Trento, Italy). A Student Research Forum (SRF) is also featured. In addition, ISVLSI 2022 included a Research Demo Session (RDS), an Embedded Tutorial, a Panel, and two after-lunch Plenary Talks by Onur Mutlu (Professor in Computer Science at ETZ Zurich) and Vijay Narayanan (A. Robert Noll Chair Professor of Computer Science & Engineering and Electrical Engineering at the Pennsylvania State University).

As we have successfully completed 20 editions of the symposium as a conference (and 4 editions where ISVLSI was run as a workshop, we have decided to honor the conference’s legacy and pay tribute to its forefathers, the late Professors Amar Mukherjee and Nagarajan “Ranga” Ranganathan. As such, the 2022 edition featured a plenary panel session with some key members of our community participating, moderated by Professor Vijay Narayanan, that highlighted the legacy of the conference so far, with the panelists sharing their viewpoints on how future editions should be shaping up to adapt to the changing conditions we are facing post-pandemic.

Theocharis Theocharides  
University of Cyprus, CYPRUS  
ttheocharides@ucy.ac.cy

Muhammad Shafique  
NYU Abu Dhabi, UAE  
muhammad.shafique@nyu.edu

Hai "Helen" Li  
Duke University, USA  
ttheocharides@ucy.ac.cy
An Efficient Accelerator of Deformable 3D Convolutional Network for Video Super-Resolution

Siyu Zhang, Wendong Mao, and Zhongfeng Wang, *Fellow, IEEE*

School of Electronic Science and Engineering, Nanjing University, P. R. China

Recently, deformable 3D convolutional network (D3D-ConvNet) is gaining popularity in video super-resolution (VSR) task, because it can align multiple frames and extract spatio-temporal information. However, the expensive computational complexity and irregular memory accesses caused by deformable 3D convolution (D3D-Conv) impede the deployment of D3D-ConvNet on edge devices. Firstly, dynamically produced offsets lead deformable convolutions to sample arbitrary respective fields, causing irregular memory access (IMA) patterns. Secondly, IMA causes adverse effects on data reuse. Consequently, the memory requirements and the possibility of access conflicts are increased significantly. Especially for some low-level vision tasks, the impact of this problem on memory resource overhead is further exacerbated. Thirdly, compared with the standard convolutional computing unit, the additional receptive field deformation stages and three-dimensional convolutional operations lead to high computational complexity. On the foundation of these understanding, an algorithm and hardware co-optimization framework is proposed to accelerate D3D-ConvNet for VSR on field-programmable gate array (FPGA) in this work [1].

At the algorithm-level optimization, a tile decoupling computing strategy (TDS) is put forward to relieve excessive memory overhead. The overall process is shown in Fig. 1. TDS aims at decoupling the input feature maps into overlapping tiles. Multiple stages of D3D-Conv, including 3D convolutions and bilinear interpolation (BLI), are executed continuously at the tile-level granularity. After that, the values in the non-overlapping positions of each tile are collected sequentially and utilized to compose the output feature maps. TDS helps us to relieve the memory resource requirements cased by the deformed receptive field and large feature map and get rid of the obvious performance degradation sufficiently.

Depending on TDS, several computing modules together with a ping-pong transposition storage scheme (PTS) and an overall hardware architecture are developed to accelerate D3D-Conv operations. Fig. 2 demonstrates the overall hardware architecture of D3D-Conv accelerator. Four dual-port SRAMs, Input buffer, Weight buffer, Offset buffer, and Output buffer, are used to store data. Top controller is responsible for scheduling the computation orders of the on-chip units. Address converter generates a series of neighboring integer addresses of the on-chip data buffer. Coefficient generator is utilized to derive BLI coefficients. Parallel vector-based dot productions are performed by a group of BLI cores. PTS is composed of two register arrays with the same size, which mainly takes charge of storing and transferring the intermediate data. 3D convolutional computing array (CCA) is utilized to perform parallel 3D convolutional operations. Accumulator, including several registers and adders, is employed to accumulate intermediate calculations. In a nutshell, the proposed design relieves excessive memory requirements caused by the deformed receptive field and large input resolution. Meanwhile, since PTS caches intermediate features and ensures them to be accessed regularly, our design overcomes access conflicts and accelerates deformable convolutional layers tremendously.

The proposed architecture is implemented on Xilinx Virtex UltraScale+ FPGA platform. In terms of algorithm visualization, the proposed method can reconstruct sharper textures and richer details. Compared with prior FPGA-based VSR methods, experimental results demonstrate more promising visual perceptions and better accuracy in PSNR and SSIM. In terms of hardware efficiency, the proposed accelerator can reach 295.29 GOPS at 200 MHz, which surpasses existing hardware implementations for deformable convolutions significantly.

References

Adaptable Multi-level Voltage to Binary Converter Using Ferroelectric FETs

Sanjay Das1, Arun Govindankutty1, Shan Deng2, Kai Ni2, Sumitha George1

1North Dakota State University, Fargo, USA
2Rochester Institute of Technology, NY, USA

There is a growing need for high density storage devices and their effective implementations due to the increase in data-intensive applications. Considering its promise of greater storage density and reduced unit storage cost, multi-level-cell(MLC) memories are being researched extensively. Nevertheless, to process the multi-bit data that is stored in these memories, it must be transformed into formats that are compatible with the processor(usually binary). HfO2 based ferroelectric field effect transistors (FeFET), for its superior energy-efficiency, great scalability, and CMOS-compatibility, has received significant attention for various logic and memory designs, whereas a few works have explored the possibility of using FeFETs for peripheral sensing purposes. Here we have suggested an adaptable multi-level voltage to binary converter using FeFETs that can convert input voltage to bits.

The proposed converter implementation is compact, as we directly utilize threshold voltage(Vt) of FeFETs for voltage comparison. The circuit is versatile as it provides several valuable features such as adaptable quantization steps, flexible output bit width and an inherent security capability (Fig. 1). As multiple domains in Fe-layer influences the number of polarization states and threshold voltages(Vt's) in a FeFET, choosing different Vt's for FeFETs offer different quantization steps for the conversion. The converter circuit can also function in different bit resolutions(2-bit, 1-bit etc.) without any explicit changes to the circuit. For instance, the 2-bit converter(Fig. 2) can function as a 1-bit converter by using only one FeFET as voltage comparator and its source as the output. This flexible output bit-width capability provides designers with freedom to trade bit density for reliability in data conversion. Additionally, the order in which FeFETs in the circuit are programmed with different Vt's influences generation of the output bits. We utilize this inherent encryption capability to introduce security in the design.

The proposed circuit is shown in Fig. 2, which converts 4-discrete level voltages to 2-bit output. It consists of three FeFETs(F1, F2, F3), three NFETs(M1, M2, M3), and a PFET(M4). FeFETs function as voltage comparator to generate thermometer codes. NFETs are used to provide a discharge path for residual voltages at FeFET source terminals. PFET works as an encoder for the output. The FeFETs(F1, F2, F3) are programmed to threshold voltages Vt1, Vt2 and Vt3 respectively(Vt1<Vt2<Vt3) by applying distinct write pulses. A pre-discharge signal is applied to M1, M2 and M3 before applying input for reliable conversion. When an input voltage(Vin) is applied, the FeFETs with Vt<Vin get switched ON and passes Vin(500mV) to the source terminals. The polarization states remain unchanged during the conversion, which is ensured by restricting the pulse amplitude and width of Vin appropriately. The selective turning ON of FeFETs generate distinct thermometer codes for various inputs. M4 at the output stage encodes the thermometer code to corresponding binary output. The source of F2 is the MSB of the output and the gate input to PFET. When F2 is switched ON, M4 is OFF and source of F3 is selected as LSB of output. When F2 is OFF, M4 is turned ON and passes source signal of F1 as LSB of the output.

Additionally, the circuit uses progressive output encoding, which reduces error margin to least significant bit(LSB). By simulating the suggested 4-level to 2-bit converter circuit with 20 and 2000 domains FeFET, we convert input voltages between [0 - 3.75V] and [0 - 2.7V] to output bits respectively. The design is scalable for more quantized input voltage levels and corresponding output bits by adding more FeFETs (to increase the number of voltage comparison states) and extending the output encoding circuit.

**References**
WOMEN-IN-VLSI (WiV) SERIES: Dr. Diana Marculescu

Dr. Marculescu is Chair of the Electrical and Computer Engineering at University of Texas in Austin

Here, she shares more about her work and the future of her field.

Q1. What VLSI/Computing research area have you focused on during your career?

My early research centered around Computer Aided Design (CAD) tools for energy efficient system design, with a later focus on reliability and variability, as well as system power management for sustainable computing. I am currently working in the area of hardware architecture for machine learning (ML), hardware-ML co-design, and hardware-aware/energy efficient ML systems.

Q2: Why do you think this area is important currently?

The power consumption of digital systems has become an important design constraint about three decades ago and has continued to be a top priority in VLSI design. With the increased importance of sustainable computing paradigms and their impact on the digital economy and environment, there is no surprise that this is and will continue to be a prime focus for next generation computing systems. I don’t see this interest going away anytime soon, in fact it will increase in impact given the importance of the digital economy on climate and environment. The students we educate or the researchers we train will need to be savvy in the impact that the systems they’re building have on the environment – this may be quite a departure from the way engineering education and research has been traditionally done.

Q3: What motivated you into this field?

I am a trained computer scientist (through my undergraduate degree) and a computer engineer (through my PhD) so when time had come to decide what area to work on as a PhD student it was quite apparent that for maximum impact one needs to understand the entire system stack and advance the field in a synergistic fashion. I had excellent mentors in both my undergraduate and graduate education, many of whom were respected women scientists, engineers, and role models. This is of utmost importance for the next generation of researchers – ensuring that they find the support and inspiration to persist and successfully enter a career in this field – and this is something I am very passionate in my day to day activities.

Q4: What is your typical day like as the Chair of the ECE Dept at UT Austin?

There is no typical day! I make sure my time as a chair is spent with students, faculty, staff, alumni, and supporters of our department, so depending on the day, I may be in retreats or research brainstorming sessions with faculty colleagues, formal or informal meetings with staff members, or student or alumni events advancing our mission. While I do have standing meetings for the staff and faculty I work directly with, I always have my door open for anyone who needs help navigating an issue or support in pushing forward a new initiative. I have an active research group of six PhD students and many other MS and undergraduate researchers, so balancing everything is something I strive for every day.

Q5: What excites you most when you look into the future?

First and foremost, the students! They are exceeding my expectations every day in what they achieve in the classroom, in the lab, and in real world after they become dedicated alumni. The future of electronics and semiconductor industry, as well as academic relevance of this field is in great hands given what I see from the youngest members of our community. Of course, the progress made in new technologies and what they can bring to the development of new computing systems is truly exciting, as they can disrupt how we design, manufacture, and manage computing resources in the future.
Q6: Many students of color and women worry that the VLSI STEM field won’t welcome them. When you look at the landscape for academia and industry, what do you see right now? Are there signs of progress?

There has been a lot of progress without a doubt, but like with many of the changes we’ve seen in history, progress isn’t linear, and not necessarily monotonic. The overall gradient from decades ago when I entered the field is definitely positive, but there is quite a bit left to do. Achieving gender parity, reflecting the demographics of population at large back in our field, coupled with robust mechanisms for including and empowering everyone to achieve their goals regardless of their upbringing, and providing equitable support to those historically marginalized are a work in progress and require everyone’s buy in, which the majority of stakeholders in both industry and academia embrace wholeheartedly.

Q7: How do you balance your work life and family life?

I have enjoyed the help of my close family, friends, and support network in balancing a hectic professional life with a normal family life, but it shouldn’t come as a surprise that many times these have to be intertwined to make things happen. Indeed, my three children – all of whom are teenagers or young adults now – have traveled with me at the many conferences I had to attend or during my sabbatical leaves. Making explicit time for everyone is crucial, and shouldn’t be neglected. And most importantly, making time for yourself is essential. A balance between work and family life isn’t possible if we don’t make the time for both, so this is what I strive for every day.

Q8: What is your key message to young girls who aspire to be like you someday?

First, don’t be like me – be like you, and always measure yourself against yourself, not others! What’s most important is to move in the right direction that works for you, so find that guiding light with the help of mentors. Second, build a support network and ensure your family is on hand to support you in your endeavors. Last (and definitely perhaps most important), prioritize and learn to distinguish what’s important from what’s urgent – suffice to say the former is what counts! Your well being should be first priority as it is what drives a healthy and successful life both personally and professionally.

Diana Marculescu is Department Chair, Cockrell Family Chair for Engineering Leadership #5, and Professor, Motorola Regents Chair in Electrical and Computer Engineering #2, at the University of Texas at Austin. Prior to joining UT Austin in December 2019, she was the David Edward Schramm Professor of Electrical and Computer Engineering, the Founding Director of the College of Engineering Center for Faculty Success (2015-2019) and has served as Associate Department Head for Academic Affairs in Electrical and Computer Engineering (2014-2018), all at Carnegie Mellon University. She received the Dipl. Ing. degree in computer science from the Polytechnic University of Bucharest, Bucharest, Romania (1991), and the Ph.D. degree in computer engineering from the University of Southern California, Los Angeles, CA (1998). Her research interests include energy- and reliability-aware computing, hardware aware machine learning, and computing for sustainability and natural science applications. Diana was a recipient of the National Science Foundation Faculty Career Award (2000-2004), the ACM SIGDA Technical Leadership Award (2003), the Carnegie Institute of Technology George Tallman Ladd Research Award (2004), and several best paper awards. She was an IEEE Circuits and Systems Society Distinguished Lecturer (2004-2005) and the Chair of the Association for Computing Machinery (ACM) Special Interest Group on Design Automation (2005-2009). Diana chaired several conferences and symposia in her area and is currently an Associate Editor for IEEE Transactions on Computers. She was selected as an ELATE Fellow (2013-2014), and is the recipient of an Australian Research Council Future Fellowship (2013-2017), the Marie R. Pistilli Women in EDA Achievement Award (2014), and the Barbara Lazarus Award from Carnegie Mellon University (2018). Diana is a Fellow of ACM and IEEE.
(1) **Biden signs $280 billion CHIPS and Science Act**

President Joe Biden signed the CHIPS and Science Act on Tuesday, writing into law the $280 billion package that includes $52 billion in funding to boost US domestic semiconductor manufacturing. This funding authorization brings Intel and other chipmakers one step closer to building out plants in states like Ohio and Arizona, projects that are reliant on the subsidies.

(2) **Latest HYPERRAM boosts throughput-per-pin**

HYPERRAM—the expansion memory for scratchpad and data buffering—has reached its third generation and is ready to serve a wide range of applications such as video buffering, factory automation, artificial intelligence (AI) edge processing, and automotive vehicle-to-everything (V2X) communications. The third generation of HYPERRAM devices from Infineon Technologies support the newly extended HyperBus interface to enable 800 MBps data rates at a density range of 64 Mb to 512 Mb.

(3) **Neuromorphic Chip Gets $1 Million in Pre-Orders**

Neuromorphic computing company GrAI Matter has $1 million in pre–orders for its GrAI VIP chip. The GrAI VIP chip is an SoC with an updated version of the company’s neuron flow fabric plus dual Arm Cortex M7 CPUs (including DSP extensions) for pre- and post–processing. It has dual MIPI Rx/Tx camera interfaces. GrAI VIP can handle MobileNetv1–SSD running at 30fps for 184 mW, around 20× the inferences per second per Watt compared to a comparable GPU, the company said, adding that further optimizations in sparsity and voltage scaling could improve this further.

(4) **Computing and Storage at the Same Time! Macronix Creates Innovative Memory**

Macronix’s FortiX series 3D flash memory supports computing-in-memory (CIM) / in-memory search (IMS) functions, which can share CPU computing tasks, improve system speed and reduce power consumption. Macronix’s FortiX IMS 3D NAND accelerator greatly reduces the data movement of the subsequent operations of a von Neumann architecture system.

(5) **GPUs Are Role-playing Quantum Computers**

Graphics processors are taking on a new role beyond gaming and artificial intelligence – they are now serving as surrogate quantum computers until the real hardware arrives. The Jülich Supercomputing Centre is using GPUs and a software toolkit from Nvidia to emulate quantum computers and research algorithms for such systems. With quantum processors still under development, GPUs are the fastest circuits to play the role of fully operational quantum computers.

(6) **GlobalFoundries unveils new silicon photonics platform**

Pure play foundry GlobalFoundries Inc. (GF) has introduced a new silicon photonics manufacturing process designed for data center solutions. Fotonix is a monolithic platform that combines GF’s differentiated 300 mm photonics features and 300 GHz-class RF-CMOS on a silicon wafer. Fotonix consolidates complete processes that were previously distributed across multiple chips by combining a photonic system, radio frequency components and CMOS onto a single silicon chip. The 300 mm monolithic silicon photonics solution uses high data rate per fiber (0.5 Tbps/fiber) to enable 1.6-3.2 Tbps optical chiplets for more efficient transmission of data and better signal integrity.
TCVLSI Sponsored Conferences for 2022

Financially sponsored/co-sponsored conferences

- ARITH, IEEE Symposium on Computer Arithmetic
- ASAP, IEEE International Conference on Application-specific Systems, Architectures and Processors
  - ASAP 2022: https://www.asap2022.org/ Virtual conference dates: July 12-14 2022
- ASYNC, IEEE International Symposium on Asynchronous Circuits and Systems
  - ASYNC 2022: https://asyncsymposium.org/async2022/ Virtual conference dates: TBD 2022
- iSES, (formerly IEEE-iNIS) IEEE International Smart Electronic Systems
  - IEEE iSES 2022: https://ieee-ises.org/2022/ Dec 21-23 2022, NIT Warangal, India
- ISVLSI, IEEE Computer Society Symposium on VLSI
  - ISVLSI 2022: http://www.eng.ucy.ac.cy/theocharides/isvlsi22/ Virtual conference dates: July 4-6 2022
- IWLS, IEEE International Workshop on Logic & Synthesis – collocated with DAC
  - IWLS 2022: https://www.iwls.org/iwls2022/ Conference dates: July 18-21, 2022
- SLIP, ACM/IEEE System Level Interconnect Prediction
  - SLIP 2022: https://dl.acm.org/conference/slip/proceedings Date TBD

Technically Co-Sponsored Conferences for 2022

- VLSID, International Conference on VLSI Design
  - VLSID 2022: https://vlsid.org/ Virtual conference dates: Feb 26 -March 2 2022

Explore conference sponsorship options with TCVLSI here: https://www.computer.org/conferences/organize-a-conference/sponsorship-options