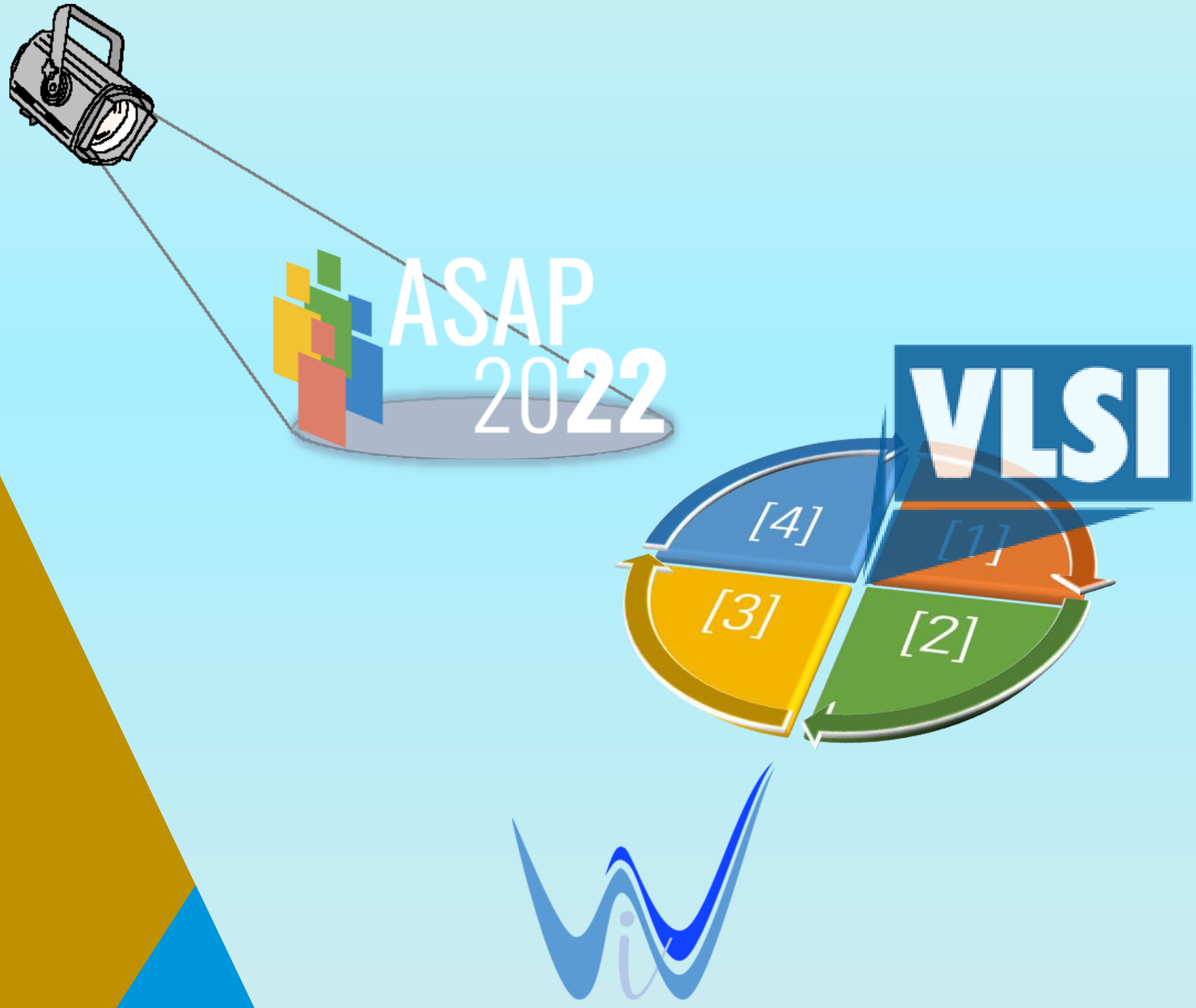
IEEE VLSI Circuits & Systems Letter

QUARTERLY PUBLICATION OF
IEEE COMPUTER SOCIETY TECHNICAL COMMITTEE ON VLSI (TCVLSI)

Volume 8 Issue 4 November 2022



Editor-in-Chief:
Mondira Pant

IEEE VLSI Circuits and Systems Letter

Volume 8, Issue 4, Nov 2022

Editorial

Invited articles

- **“Experiential VLSI for Undergraduate Curriculum**” - *Prof Arijit Raychowdhury (Steve W Chaddick Chair of the School of Electrical and Computer Engineering, Georgia Institute of Technology)*

Conference spotlight

- ASAP 2022 – Report by Program Co-Chairs – *Ioannis Sourdis, Miquel Pericas and Dionisios N Pnevmatikatos*

- ASAP 2022 Best Paper “Answer Fast: Accelerating BERT on the Tensor Streaming Processor” - *Ibrahim Ahmed, Sahil Parmar, Matthew Boyd, Michael Beidler, Kris Kang, Bill Liu, Kyle Roach, John Kim, Dennis Abts*

- ASAP-2022 Best Paper Nominee “HPVM2FPGA: Enabling True Hardware-Agnostic FPGA Programming”- *Adel Ejjeh, Leon Medvinsky, Aaron Councilman, Hemang Nehra, Suraj Sharma, Vikram Adve, Luigi Nardi, Eriko Nurvitadhi, Rob A Rutenbar*

- ASAP-2022 Best Paper Nominee: “LOSTIN: Logic Optimization via Spatio-Temporal Information with Hybrid Graph Models” - *N. Wu, J. Lee, Y. Xie and C. Hao*

- ASAP-2022 Best Paper Nominee “Efficient Synchronization for NR-REDCAP Implemented on a Vector DSP” – *S. F. Qureshi, S. A. Damjancevic, E. Matus, D. Utyansky, P. Van der Wolf and G. P. Fettweis*

Women in VLSI (WiV) series spotlight

- Interview with Prof Sandhya Dwarkadas, Walter N. Munster Professor and Chair of Computer Science Department at University of Virginia

Updates

- Recent relevant news highlights – *Ishan Thakkar*
- 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 Societ**y. 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.

IEEE TCVLSI 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 an invited article “***Experiential VLSI for Undergraduate Curriculum***” by Prof Arijit Raychowdhury Dept Chair of ECE at Georgia Tech, where while discussing the challenges, underscores the importance of an undergraduate curricular structure that provides students hands-on experience in designing, taping-out and measuring silicon as part of the degree requirement.

The newsletter spotlights one of TCVLSI’s sponsored symposiums in 2022 IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP). One-page teasers of the best paper awarded at ASAP 2022 and three best paper nominees are showcased: ***“Answer Fast: Accelerating BERT on the Tensor Streaming Processor"; “HPVM2FPGA: Enabling True Hardware-Agnostic FPGA Programming”; “LOSTIN: Logic Optimization via Spatio-Temporal Information with Hybrid Graph Models”; “Efficient Synchronization for NR-REDCAP Implemented on a Vector DSP”.***

In our Women in VLSI (WiV) series, we share an inspiring interview with Prof. Sandhya Dwarkadas, Walter N. Munster Professor and Chair of Computer Science Department at University of Virginia

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 Nov 2022 and a newletter readership of about 30,000
To join TCVLSI (its free), click here: https://www.ieee.org/membership-catalog/productdetail/showProductDetailPage.html?product=CMYVLSI732

Experiential VLSI for the Undergraduate Curriculum

Arijit Raychowdhury
School of Electrical and Computer Engineering
Georgia Institute of Technology

Experiential learning is a critical component of the undergraduate curriculum in most of our universities. In the Electrical and Computer Engineering programs, students engage in hands-on, experimental work either as a part of their instructional classes or as separate laboratory classes. This model has been extensively successful across the world in areas as varied as signal processing, embedded systems, computing, power systems, circuits as well as in emerging areas such as machine learning and artificial intelligence. The emergence of Makerspaces on college campuses has provided students with unique opportunities to conceptualize engineering designs, develop prototypes, validate them, test them and finally using the learnings to refine and re-engineer the systems, thus closing a critical loop in the design cycle. In spite of the great success of this experiential learning model some areas of Electrical and Computer Engineering have not been able to incorporate this philosophy in their classes. VLSI is one such area where the demand for practical design experiences has been hard to meet. Before looking at the challenges of experiential VLSI learning, let us explore the current need and the opportunities.

The world of semiconductors is at a critical juncture. The severe supply chain crisis in the area of semiconductor devices, chips and systems has been exacerbated over the last few years by the pandemic. This has highlighted the need for growth in the area of semiconductors, from device fabrication and manufacturing, design, packaging and system integration. We are at a crossroad where we are not only experiencing a shortage in the market, but also noticing severe gaps in the talent pipeline. This has become a national as well as a global crisis and the significant investments from governments across the globe are being made. This is augmented by investments by semiconducting companies that are interested in not only foundational and applied research, but also training the next generation of the workforce in this highly skilled area. In the context of

VLSI design, it is enticing to think of a curricular structure for our undergraduate students where they get hands-on experience in designing, taping-out and measuring silicon as a part of their degree requirement. At the School of Electrical and Computer Engineering at Georgia Tech, we have recently introduced a new undergraduate course with the objective of providing students in the VLSI Design curricular thread to experience the thrills, challenges and satisfaction of prototyping their first digital SoC in a scaled foundry CMOS process within the confines of a classroom.

Although the premise is promising, several practical challenges remain in introducing experiential VLSI in an undergraduate class. Firstly, a complete design experience in a programmable, digital SoC spanning across RTL design, front-end and back-end designs, power-signal integrity checks, timing closure, and top-level system integration is considerably challenging for undergraduate students. This necessitates that we setup a pipeline of classes starting from the sophomore year that will train the students about circuit theory, computer architecture, design modularity, CAD tools, design verification and testing, that will culminate into a senior class where the students will go through the process of silicon design and tape-out. This requires restructuring, and sometimes revamping our course offerings and allocating enough resources to provide coherent, multi-year training in the art of VLSI design. Secondly, a single semester may be enough for a cohort of highly motivated students to complete an end-to-end design and tape-out at the end of the semester. However, a quarter system does not allow that. Further, it takes a foundry months to manufacture the test chips, which makes testing within the timeframe of a class more challenging. This fundamentally requires course structures and course offerings that span across semester and quarter boundaries, which are often hard to implement in the undergraduate curriculum. Thirdly, negotiating the “Non Disclosure Agreements” with foundries to allow undergraduate students access to scaled CMOS nodes requires institutional support. Finally, such a learning experience is expensive to offer and considerable support from corporate partners, alumni and donors or government agencies is required. The last challenge is particularly daunting for the area of VLSI where the costs associated with taping-out and measuring silicon chips are significant, increasing and recurrent.

In spite of the challenges, we strongly believe that this is a societal need of the hour and the future. We need talented circuit designers in the semiconductor engineering community. At the same time, the barrier-to-entry in this area is significantly higher than in other areas of computing where the impact of one's work is easy to instantaneously perceive. On the other hand, VLSI design requires discipline, diverse skill sets, foundational knowledge, team-work, perseverance, and dedication over a significant period of time. But the end result — being able to program a chip that one has designed and taped-out — is immeasurably satisfying.

As I had mentioned, at Georgia Tech we have offered the first undergraduate VLSI class where the students are using the TSMC 65nm GP CMOS process to tape-out a RISC-V processor with custom math accelerators on an embedded bus, accessible and programmable through a standard ISA. We thank Apple for their support, sponsorship as well as engagement with the students through their New Silicon Initiative. Responses from the students have been overwhelmingly positive and our enrollment in VLSI classes have more than doubled in the last three years. The hands-on experience is a major driver for connecting theory to practice; thus creating more opportunities for inclusion of female and underrepresented minority students in the VLSI community. As we refine our program and learn from the current offerings, we believe that such unique experiences offered in our classrooms will attract and retain top and diverse students in circuit design. In turn, these students will fuel the next fifty years of growth and prosperity in a dynamic industry where we all belong.

Arijit Raychowdhury is the Steve W Chaddick Chair of the School of Electrical and Computer Engineering at the Georgia Institute of Technology. He was previously the Motorola Foundation Professor in the School. From 2013 to July 2019 he was an Associate Professor and held the ON Semiconductor Junior Professorship. His industry experience includes five years as Staff Scientist with the Circuits Research Lab, Intel Corporation, and two years as an Analog Circuit Researcher with Texas Instruments Inc. Dr. Raychowdhury's research interests include low-power digital and mixed-signal circuit design and exploring interactions of circuits with device technologies. He has authored over 200 articles in journals and refereed conferences and holds more than 26 U.S. and international patents. His recent awards include the 2021 SRC Technical Excellence Award, the 2020 Qualcomm Faculty Award and 2018 IEEE/ACM Innovator under 40 Award. He is a Fellow of IEEE and currently serves on the technical program committees of ISSCC, CICC, VLSI Symposium and DAC. He is an IEEE SSCS Distinguished Lecturer for 2021-2022.

ASAP 2022

33rd IEEE International Conference on Application-specific Systems, Architectures and Processors

The ASAP conference has been running for over three decades, but its technical focus seems today more relevant than ever. That is because the benefits of traditional technology scaling, namely Dennard Scaling and Moore's Law, have been reducing significantly and Application- and Domain-specific solutions are a promising way to continue improving computing performance and energy-efficiency previously offered by technology scaling.

This year's program included, as detailed next, five regular sessions on the topics of tools, parallel processing, machine learning, reconfigurable systems, application-specific processing and security, a special session on European acceleration technologies, and four keynote talks.

The first session focused on *Tools* for efficient system generation, covering tools for hardware-agnostic FPGA programming, logic optimization with hybrid graph models, the memoryless synthesis of weightless neural networks, and the optimization of near- or on-chip memory computing systems. The *Efficient Parallel Processing* session had an emphasis on accelerators, covering vector processor design, efficient processing on vector DSP and GPUs, and techniques to exploit Processing-in-Memory potential using Hugepages. The *Machine Learning* session covered a wide range of topics from performance, addressing latency with DNN task mapping and improving the BERT performance on Tensor Streaming Processors, core processing efficiency proposing a low-precision logarithmic arithmetic for neural network accelerators, and a systolic multiplication architecture for graph convolutional networks, and the application of deep learning for intrusion detection in automotive networks CAN bus. The *Reconfigurable systems* session covered implementation issues for deep-learning addressing accelerators both for inference for Binary and Quantised Neural Network Accelerators, as well as learning using quantised convolution neural networks. The session also included work on a path planning accelerator for autonomous driving using hardware-assisted memorization. The *Application specific processing and Security* session covered works on efficient object detection and pattern recognition, the codesign of Application-Specific Instruction-Set RISC-V processors, and the efficient verification for permissioned blockchains and secure communication for NoCs. Finally, the special session included invited talks on *Acceleration Technologies developed in Europe*, covering advances in reconfigurable event-triggered computing, embedded FPGAs for the future high-performance processors and RISC-V specific Systolic Array, Vector, and AI Accelerators.

The first keynote was given by Christos Kozyrakis from Stanford University and focused on Systems Support for Accelerators. Christos described the challenges in building large-scale accelerated systems: feeding accelerators with data (data ingestion and storage), building memory systems for large datasets with irregular accesses, and providing security and privacy guarantees for accelerated tasks. He presented encouraging first results on these three topics and suggested open questions that our community needs to address in future research.

Jan Andersson from Cobham Gaisler AB discussed the state of the art, approaches and trends (including open-source hardware) in designing microprocessors for space applications. Sharing experience from the long-running Gaisler's operations that range from architectural design of microprocessors to radiation- and environmental-testing of systems, Jan motivated the move to higher performance (multi-core and accelerated) in the space application domain so as to increase the on-board (pre-)processing and mitigate the downlink capacity limitation impact. Jan also discussed the NOEL-V RISC-V processor, Gaisler's latest major addition to its open hardware library.

Maria Girone, the Chief Technology Officer from CERN openlab, discussed the unprecedented computing challenges faced in the high-energy physics experiments done at CERN, and highlighted the increasing gap between the increase in the data collection rate stemming from the complexity of particle-collision events as compared to the the gains expected from technology evolution. This gap needs to be addressed with efficient use of HPC, new AI/ML solutions and heterogeneous architectures. Maria also presented the CERN work in software development for heterogeneous architectures, data management at scale, and supporting services and identified open problems that pose future opportunities for R&D.

The last keynote was offered by Alex Ramirez, a Sr. Staff at Google discussing the dos and do-nots in designing accelerators. Alex used his journey, from concept to planet-scale deployment, of the Video transCoding Unit - that Google developed to accelerate the video processing at YouTube- to derive lessons and suggestions for the successful co-design and deployment of future hardware accelerator platforms. A few of the takeaways include having an engaged customer from day 0, focusing on software readiness, checking the validity of assumptions along the path, and never underestimating the impact of the things that were not planned-ahead for.

ASAP 2022

33rd IEEE International Conference on Application-specific Systems, Architectures and Processors

IEEE ASAP 2022 best papers:

Five papers were selected from the IEEE ASAP program as the best of this year [1][2][3][4][5]. Their focus is on tools and architectures for application specific processing and target various domains including machine learning. A brief description of these five papers follows next:

The HPVM2FPGA paper describes a novel end-to-end compiler and auto-tuning system that can automatically tune hardware-agnostic programs for FPGAs [1]. It uses a hardware-agnostic abstraction of parallelism as an intermediate representation (IR) to represent hardware-agnostic programs. HPVM2FPGA's powerful optimization framework uses sophisticated compiler optimizations and design space exploration to automatically tune a hardware-agnostic program for a given FPGA and offers up to 33x speedup.

LOSTIN proposes a novel approach to overcome the stringent accuracy requirements and the generalization capability of machine learning [2]. It aims at high quality-of-results (QoR) for ML-based logic synthesis using hybrid graph neural networks (GNNs) for accurate QoR estimations. The key idea is to simultaneously leverage spatio-temporal information from hardware designs and logic synthesis flows to forecast performance (i.e., delay/area) of various synthesis flows on different designs. Evaluation on 3.3 million data points shows that LOSTIN reduces the testing mean absolute percentage error (MAPE) of designs by 7-15x compared to previous work.

LogicWiSARD targets Weightless Neural Networks (WNNs), an alternative pattern recognition technique where RAM nodes function as neurons [3]. LogicWiSARD creates compressed minimized implementations converting trained WNN nodes from lookup tables to logic functions and reduces energy consumption by more than 80% compared with a multilayer perceptron network (MLP) and by 32.2% and 99.6% compared to other FPGA-based WNNs, convolutional neural networks, and binary neural networks.

The paper titled "Efficient Synchronization for NR-REDCAP Implemented on a Vector DSP" focuses on NR-REDCAP, the new 3GPP standard for low-cost 5G IoT devices, and in particular on the computationally intensive task of keeping such devices synchronized with the 5G network. Acceleration of synchronization algorithms on vector architectures is challenging because of data dependencies in the recursive filter that is usually employed. The paper describes how to vectorize the synchronization algorithm, including its recursive filter, on a wide SIMD vector DSP and achieves a speedup of 9x over scalar processing.

The “Answer Fast” paper focuses on transformers workloads for machine learning and in particular for the ones used in real time [5]. Transformer computations include, besides matrix multiplications, several non-linear components, which tend to become the bottleneck during an inference. Answer Fast describes an accelerator for the inference of BERT models on the tensor streaming processor. It fuses all non-linear components with the matrix multiplication components utilizing the on-chip matrix multiplication units more efficiently. In turn, this results in a deterministic tail latency, which is 6x faster than the current state-of-the-art.

Ultimately, a committee composed of IEEE ASAP 2022 TPC members selected the “Answer Fast” paper [5] as the best paper of this year’s program. Next are listed one pagers on the “Answer Fast”, “HVPM2FPGA”. “LOSTIN” and “Efficient Synchronization for NR-REDCAP Implemented on a Vector DSP”.

[1] Adel Ejjeh, Leon Medvinsky, Aaron Councilman, Hemang Nehra, Suraj Sharma, Vikram Adve, Luigi Nardi, Eriko Nurvitadhi, Rob A Rutenbar, "HPVM2FPGA: Enabling True Hardware-Agnostic FPGA Programming"

[2] N. Wu, J. Lee, Y. Xie and C. Hao, "LOSTIN: Logic Optimization via Spatio-Temporal Information with Hybrid Graph Models"

[3] Igor D.S. Miranda, Aman Arora, Zachary Susskind, Luis A.Q. Villon, Rafael F. Katopodis, Diego L.C. Dutra, Leandro S. De Araújo, Priscila M.V. Lima, Felipe M.G. França, Lizy K. John, Mauricio Breternitz, "LogicWiSARD: Memoryless Synthesis of Weightless Neural Networks,"

[4] S. F. Qureshi, S. A. Damjancevic, E. Matus, D. Utyansky, P. Van der Wolf and G. P. Fettweis, "Efficient Synchronization for NR-REDCAP Implemented on a Vector DSP"

[5] Ibrahim Ahmed, Sahil Parmar, Matthew Boyd, Michael Beidler, Kris Kang, Bill Liu, Kyle Roach, John Kim, Dennis Abts, "Answer Fast: Accelerating BERT on the Tensor Streaming Processor"

Answer Fast: Accelerating BERT on the Tensor Streaming Processor

Ibrahim Ahmed, Sahil Parmar, Matthew Boyd, Michael Beidler,
Kris Kang, Bill Liu, Kyle Roach, John Kim and Dennis Abts
Groq Inc.
{iahmed, sparmar, matt, mb, kkang, bliu, kroach, jkim, dabts}@groq.com

Transformer-based machine learning models have revolutionized various natural language processing (NLP) applications; state-of-the-art results in machine translation, web search, question and answering almost exclusively use transformers. Many of the production-level transformer-based models are real-time systems where users interact with a service and expect a response in real-time, which enforces a strict latency requirement during inference.

BERT is a popular transformer model that is widely used in the industry: Microsoft and Google search engines rely on BERT models; and Twitter content moderation pipeline also includes a BERT model. In many services, an inference through the BERT model is usually one component that feeds other downstream tasks before returning an answer to the user. As such, to guarantee a reasonable service time it is important to ensure that the observed latency of the inference is under the strict latency budget available to the entire pipeline. In addition, any reduction in the inference latency would relax the timing constraints for the other components in the rest of the pipeline.

The goal of this work is to accelerate inferences through BERT models with the objective of minimizing both *average* latency and latency *variation* (including tail latency). We exploit the Groq Tensor Streaming Processor (TSP) hardware accelerator that provides high batch-1 performance while supporting deterministic execution to minimize latency variation. The TSP, a statically scheduled SIMD architecture, performs computation using a streaming processing model where computational elements, arranged spatially by function, perform operations as tensor data streams over the functional units. The high-level microarchitecture and spatial layout of the TSP architecture is shown in Fig. 1. Matrix execution modules (MXMs) perform matrix-matrix multiplications, Switching execution modules (SXMs) perform data transformations, and the Vector execution module (VXM) performs point-wise vector arithmetic. SRAM memory modules (MEM) are located between the computational units providing localized, high bandwidth, low latency, load and store.

The computation involved in a BERT inference is mainly dominated by general matrix multiplications (GEMMs) which makes it amenable to acceleration on chips with dedicated matrix multiplication units. Unfortunately, the presence of the various non-linear components (softmax, layernorm and GELU) usually results in under utilizing the matrix multiplication units as they have to stay idle waiting for results from these layers. To accelerate BERT on the TSP, we leveraged the chaining capability of the VXM to pipeline the nonlinear computations with matrix multiplications such that we maximize the utilization of the MXM units on the TSP; Fig. 2 shows how we efficiently mapped one of the compute blocks (self-attention) in BERT to the TSP. By pipelining the data transformation (reorder operation) with the execution of the batched-GEMM, we can start the batched-GEMM without waiting for the reorder operation to finish execution. We were also able to hide the latency of the softmax operation completely by starting its execution as soon as the first vector of results is produced by the upstream batched-GEMM operation, and overlapping the last part of the softmax with another independent GEMM.

Our work shows that predictable performance can be achieved, compared to a modern GPU, while also achieving significantly lower average latency. For BERT-base inferences, our design results in a deterministic tail latency of 130 µ, which is 6× faster than the current state-of-the-art.

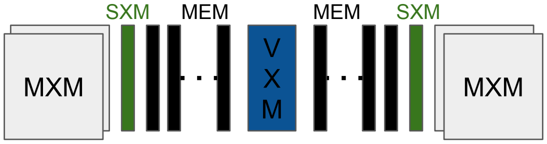


Fig. 1. Simplified view of parts of the TSP architecture.

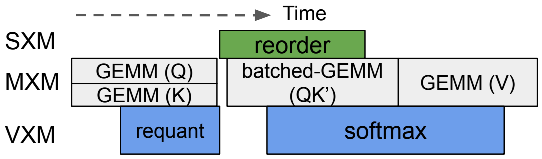


Fig. 2. Self-attention block execution schedule. Grey, blue and green are performed on MXM, VXM and SXM, respectively.

I. Ahmed *et al., ”Answer Fast: Accelerating BERT on the Tensor Streaming Processor,” 2022 IEEE 33rd International Conference on Application-specific Systems, Architectures and Processors (ASAP). Won Best paper of ASAP 2022.*

HPVM2FPGA: Enabling True Hardware-Agnostic FPGA Programming

Adel Ejjeh*, Leon Medvinsky*, Aaron Councilman*, Hemang Nehra*, Suraj Sharma*, Vikram Adve*,
Luigi Nardi†‡, Eriko Nurvitadhi§, and Rob A Rutenbar¶.
**University of Illinois at Urbana-Champaign,* Urbana, IL, USA. †*Lund University,* Lund, Sweden. ‡*Stanford University,* Stanford, CA, USA. §*Intel Corp,* Hillsboro, OR, USA. ¶*University of Pittsburgh,* Pittsburgh, PA, USA.

Recently, FPGAs have become widely available in heterogeneous systems and public clouds, taking them beyond the traditional audience of hardware designers and making them accessible to the much larger category of software application developers. However current FPGA programming paradigms are not well suited for this type of developers. Software teams deal with large complex applications, with stringent constraints on development cost, source code portability, code reuse, software security, and rapid development cycles, in addition to performance. These constraints make it difficult to invest extensive time and effort to tune application components for specific hardware targets. In many cases, this translates to an *acceptable trade-off of raw performance for improved programmability*. As such, software designers would greatly benefit from an end-to-end compiler framework that supports **hardware-agnostic programming** of FPGAs.

It is beyond doubt that hardware-agnostic programming of FPGAs *remains a holy grail* in the FPGA community. While we acknowledge the difficulty of this problem, we believe that it is achievable with the following requirements: a) an **end-to-end compiler and autotuning system** that tunes *hardware-agnostic* kernels by automatically selecting FPGA-specific optimizations, and transparently handles host code generation, b) a compiler intermediate representation that captures different kinds of parallelism (task, data, pipeline) while identifying units of acceleration, and c) a runtime system that transparently interfaces between host and device. This kind of end-to-end flow is largely missing in the FPGA design community, and state-of-the-art tools today lack one or more of these components.

We propose HPVM2FPGA, a novel *end-to-end compiler and autotuning system* that can automatically tune hardware-agnostic programs for FPGAs. HPVM2FPGA uses a hardware-agnostic abstraction of parallelism as a compiler intermediate representation (IR), building on the HPVM compiler IR [1], an explicitly parallel extension of LLVM IR designed for heterogeneous parallel systems. HPVM2FPGA adds a powerful optimization framework that uses sophisticated, parameterized, compiler optimizations (both FPGA-specific and generic) and design space exploration (DSE) to automatically tune a hardware-agnostic program for a target FPGA. It also adds a code-generation back-end that generates optimized OpenCL code for the FPGA kernels and uses the Intel FPGA SDK for OpenCL (AOC) to synthesize them.

The optimization framework includes a variety of compiler optimizations designed to automate some of the manual tuning that is required for FPGAs. These are a combination of loop-level, memory-level, and multi-kernel HPVM DFG optimizations. We use HyperMapper [2] as our DSE engine since it has been shown to optimize hardware design problems. To make DSE practical for FPGAs, we developed an analytical performance model that estimates the execution time of the optimized input program on the target FPGA, using a loop pipeline latency calculation and a critical path analysis to account for inter-kernel parallelism. This performance model is independent of specific optimizations, and derives its required inputs from a static analysis of the HPVM IR (post-optimizations) and performance metrics generated by the (fast) RTL generation stage of AOC. Moreover, our optimization framework is modular and extensible, so that more optimizations can be added relatively easily, and other DSE engines or performance models can be plugged in with minimal effort.

We performed an experimental evaluation of our compiler, which shows that our framework: a) can optimize hardware-agnostic, multi-kernel benchmarks achieving up to 33× speedup on an Arria 10 GX FPGA compared to unoptimized baselines, and b) can match hand-tuned FPGA designs in three out of four of our benchmarks.

With these features, HPVM2FPGA is meant to be a modular and extensible framework for the research community that can act as a basis for hardware-agnostic FPGA programming research, and that would keep on improving with more compiler optimizations, back end code-generation techniques, and performance estimation models/DSE. As it matures, HPVM2FPGA's goal is to support software programmers effectively by completely eliminating the need for hardware-specific details from the code, and shifting the burden of performing host-device glue code generation, and hardware-specific optimizations to the compiler and DSE.

References

[1] M. Kotsifakou, P. Srivastava, M. D. Sinclair, R. Komuravelli, V. Adve, and S. Adve, "HPVM: heterogeneous parallel virtual machine," in *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '18. New York, NY, USA: ACM, 2018.

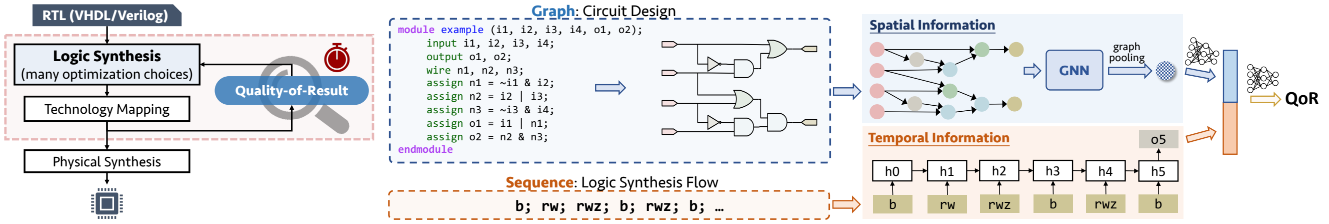
[2] L. Nardi, D. Koeplinger, and K. Olukotun, "Practical design space exploration," in *International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*. IEEE, 2019.

LOSTIN: Logic Optimization via Spatio-Temporal Information with Hybrid Graph Models

Nan Wu[1], Jiwon Lee[2], Yuan Xie[1], and Cong Hao[2]
[1]University of California - Santa Barbara, CA, USA; [2]Georgia Institute of Technology, GA, USA
{nanwu, yuanxie}@ucsb.edu, {jlee3251, callie.hao}@gatech.edu

Despite the great advance achieved by electronic design automation (EDA) tools, there is still a long way toward *hardware agile development*, whose ultimate goal is to reduce chip development cycles from years to months or even weeks. To enable rapid optimization-evaluation iterations, one mainstay is to evaluate quality-of-results (QoRs) quickly and accurately. With the emergence of machine learning (ML)-based fast performance modeling, industrial investigations highlight two basic requirements for production-ready ML applications in EDA: ① the *accuracy* of ML-based performance estimation should be a minimum of 2σ ($\sim$95%); ② the *generalization* capability to new designs is important.

In logic synthesis, hardware designs are converted to logic networks, which are typically graph abstractions of logic circuit implementations in the gate level. Logic optimization aims to transform logic networks to reduce the amount of required hardware or the critical path delay by sequences of logic transformations (i.e., logic synthesis flows). The challenges of efficient logic optimization come from two aspects. First, the design space of possible synthesis flows is extremely large, re-emphasizing the importance of fast and accurate QoR estimations (i.e., logic delay/area) to enable sufficient design space exploration. Second, existing ML-based predictive models have limited generalization capability, since they do not generalize across designs nor variable length flows. In response, we propose a novel and multi-modal learning-based approach to cope with such scenario as well as to fulfill the two fundamental requirements of production-ready ML in EDA, namely LOSTIN[1], which simultaneously leverages spatio-temporal information from circuit designs and logic synthesis flows to provide fast, accurate, and generalizable QoR estimations. As shown in Fig. 1, the structural characteristics of circuit designs naturally represented in graph format are distilled by graph neural networks (GNNs); the temporal knowledge (i.e., the relative ordering of logic transformations) in synthesis flows is extracted by long short-term memory (LSTM) networks. The separately learned graph embedding and sequence embedding have better expressiveness on each source of input modality (i.e., circuits in graphs and synthesis flows in sequences), improve the representation power and efficiency of information fusion, and significantly reduce the learning complexity and memory overhead.



```
module example (i1, i2, i3, i4, o1, o2);
    input i1, i2, i3, i4;
    output o1, o2;
    wire n1, n2, n3;
    assign n1 = ~i1 & i2;
    assign n2 = i2 | i3;
    assign n3 = ~i3 & i4;
    assign o1 = i1 | n1;
    assign o2 = n2 & n3;
endmodule
```

```
b; rw; rwz; b; rwz; b; …
```

(a) Focus of This Paper: QoR Prediction (b) Circuit Designs as Graphs + Logic Synthesis Flows as Sequences (c) Multi-Modal Graph Learning with Spatio-Temporal Information

Fig. 1. Overview of LOSTIN. (a) The focus of this paper is to accelerate the evaluation phase in logic optimization. (b) Multiple input modalities: circuit designs as graphs and logic synthesis flows as sequences. (c) Multi-modal graph learning: GNN to extract structural properties from circuit graphs (i.e., spatial information) and LSTM to extract relative ordering of logic optimizations (i.e., temporal information).

We construct a dataset based on EPFL benchmarks [1] and the logic synthesis tool ABC [2]. Evaluations on designs *seen and unseen* during training demonstrate the remarkable generalization capability of LOSTIN. For designs seen during training, the achieved mean absolute percentage error (MAPE) is less than 1.2%, 7× lower than existing ML-based approaches; for designs unseen during training, the achieved MAPE is still below 3.2%, 14× lower than existing ML-based approaches.

Looking forward, we notice two potential directions that may push LOSTIN forward. Regarding scalability to large designs, gate-level graph representations will explode with increasing design complexity and scale, necessitating hierarchical and multi-level abstractions from circuits to guarantee reasonable compute/memory costs. Regarding generalization to new logic transformations, some out-of-vocabulary techniques in natural language processing (NLP) can be adopted to promote the generalization capability to a new dimension. We expect multi-modal graph representation learning, which integrates the knowledge from other learning schemes with the conventional graph representation learning, to provide more versatility for different EDA tasks. To facilitate the interdisciplinary research, LOSTIN is open-source at https://github.com/lydiawunan/LOSTIN.

References

[1] L. Amarú *et al.*, “The epfl combinational benchmark suite,” in *Proceedings of the 24th Int. Workshop on Logic & Synthesis (IWLS)*, 2015.
[2] R. Brayton and A. Mishchenko, “Abc: An academic industrial-strength verification tool,” in *Int. Conf. on Computer Aided Verification*, 2010.

[1] Best paper candidate for ASAP 2022

Efficient Synchronization for NR-REDCAP Implemented on a Vector DSP

Sheikh Faizan Qureshi
Vodafone Chair
TU Dresden
Dresden, Germany
sheikh_faizan.qureshi@tu-dresden.de

Stefan A. Damjancevic
Vodafone Chair
TU Dresden
Dresden, Germany
stefan.damjancevic@tu-dresden.de

Emil Matus
Vodafone Chair
TU Dresden
Dresden, Germany
emil.matus@ifn.et.tu-dresden.de

Dmitry Utyansky
Synopsys, Inc.
Yerevan, Armenia
dmitry.utyansky@synopsys.com

Pieter van der Wolf
Synopsys, Inc.
Eindhoven, The Netherlands
pieter.vanderwolf@synopsys.com

Gerhard P. Fettweis
Vodafone Chair, TU Dresden
Barkhausen Institut
CeTI, TU Dresden
Dresden, Germany
gerhard.fettweis@tu-dresden.de

A surge of reduced capability and low-cost devices under the 5G framework for many Internet of Things (IoT) applications is foreseen. The new 3GPP standard, NR REDCAP [1], provides the means to develop such devices. However, keeping these inexpensive devices synchronized with the 5G network requires continuous synchronization algorithms running parallel with the receiver processing. Our analysis and literature agree that the acceleration of these algorithms on Single Instruction Multiple Data (SIMD) architectures is challenging because of the data dependencies in the recursive filter that is usually employed in synchronization algorithms and hence, requires a significant computing budget. The consequent need for activating extra hardware to perform continuous synchronization leads to rising device costs and higher power consumption which is undesirable. In this work, we address this problem by proposing an implementation that efficiently vectorizes the synchronization algorithm, including the recursive filter, on a wide SIMD vector DSP.

We consider a scenario where the user equipment (UE) already employs a VLIW SIMD vector processor (vDSP) to accommodate baseband DSP kernels like channel estimation, channel equalization, waveform demodulation, etc. [2]. To incorporate the 'always active' mode of REDCAP synchronization kernel, the vDSP offers a little remaining budget of a few tens of MHz. Hence, accommodating continuous synchronization on this small fraction of the DSP MHz budget would prevent activating extra hardware to serve the purpose.

The major limitation for an efficient VLIW SIMD implementation of the 'always active' mode is the recursive filter in the frequency offset estimator needed for received signal frequency correction. In this paper, we employ the '*recursive double algorithm*' [3] for SIMD vectorization of the recursive filter and solve the problem in the context of a more extensive system of multiple interacting blocks - subkernels within synchronization.

Although the 'recursive double algorithm' enables SIMD vectorization of recursive filter algorithms, the implementation requires data rearrangement, shuffle instructions and overhead functions leading to a significant number of stall cycles (up to 22%) dampening the achievable SIMD gain. Therefore, we have analyzed the program and data flow of the vectorized recursive filter and the whole 'always active' synchronization kernel. As an outcome of the analysis, we have reordered and interleaved the execution of code sections between subkernels into an efficient SW pipeline along with additional improvements to enhance the utilization of the VLIW slots. The proposed solution reduces the overall needed number of cycles per sample processed and at the same time increases the average number of instructions performed per cycle.

The results show that the proposed implementation achieves a speedup of 9X over scalar processing. Subsequently, the synchronization kernel can run on a limited MHz budget of the vDSP in parallel with other receiver processing kernels. Employing algorithmic and low-level optimizations like software pipelining on sub-kernel level can overcome inefficiencies resulting from high latency of data re-arrangement performed by the shuffle instruction. Resulting vDSP VLIW slots utilization is reasonably high. The speed-up attained by the SIMD implementation enables the kernel to run on a small fraction of the vDSP cycle budget.

References

[1] 3GPP, "Study on support of reduced capability NR devices," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.875, 12 2020, version 1.0.0.

[2] S. A. Damjancevic, E. Matus, D. Utyansky, P. van der Wolf, and G. P. Fettweis, "Channel Estimation for Advanced 5G/6G Use Cases on a Vector Digital Signal Processor," *IEEE Open Journal of Circuits and Systems*, vol. 2, pp. 265–277, 2021.

[3] P. M. Kogge and H. S. Stone, "A Parallel Algorithm for the Efficient Solution of a General Class of Recurrence Equations," *IEEE Transactions on Computers*, vol. C-22, no. 8, pp. 786–793, 1973.

©2021-2022 33^{rd} IEEE ASAP best paper award candidate.

<u>WOMEN-IN-VLSI (WiV) SERIES:</u> Dr. Sandhya Dwarkadas

Dr. Dwarkadas is the Walter N. Munster Professor and Chair of Computer Science Department at University of Virginia

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

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

My research is in the areas of computer architecture and experimental systems, primarily focused on parallel and distributed systems. I work at the interface of hardware and software, sometimes looking at technology influences on the architecture, and other times at how to design the software layers that interact with the hardware to improve performance, scale, and usability. Although not exclusively, in large part, my research focuses on coordination, communication, and sharing – the design of scalable (in terms of data size and core/accelerator count) coherence protocols and data movement mechanisms at the architecture level; the interaction of hardware and software, for example, through the development of policies to use the hardware mechanisms at the software level and the use of hardware performance counters for resource allocation at the operating system level; and the design of scalable runtime systems that allow parallel applications to scale to larger core counts and data sizes.

Q2: Why do you think this area is important currently and what motivated you into this field?

As a computer architect, knowledge of both application and runtime layers and underlying circuit and technology properties and changes is necessary to innovate. Despite decades of research in this area, new challenges continue to arise, driven by new applications and usage patterns at the software levels, new technologies with vastly different properties at the hardware level, and the unprecedented scale and availability of data. Properties such as time, energy, accuracy, reliability, security, and persistence require re-architecting to adapt to technology changes. Keeping up with the changing landscape is both an opportunity and a challenge that excites and motivates me.

Q3: What is your typical day like as the Chair of the CS Dept at Univ of Virginia?

Having been chair at the University of Rochester, I came in feeling like I had some idea of the demands of the position. Yet, in the space of 3 months, there has been no shortage of new challenges and crises. I am not sure there is a typical day! In addition to a chair's administrative responsibilities, I am also teaching a course on research perspectives to our first-year graduate students, and meeting with my graduate students regarding our own research. There is much to learn about the university, school, and department, and significant projected growth. I look forward to helping the department reach new heights in meeting its educational, research, and societal impact goals.

Q4: 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?

I do believe there has been good progress toward a more inclusive STEM and research community and much to be thankful for. That said, there is clearly work to be done. At universities, making faculty, staff, and students aware of best practices in inclusive interaction goes a long way in improving the environment. Companies, similarly, need to consider such training. The percentage of undergraduate women in computer science has certainly grown sufficiently to alleviate feelings of isolation, but the same is not yet true for other underrepresented populations. Helping foster a sense of belonging to the community makes all the difference.

Q5: What is your key message to the younger generation who aspire to be like you someday?

I would say, be themselves 😊. Seize every opportunity to learn new skills and topics outside of their core interests. So much of the innovation and exciting challenges is interdisciplinary that having knowledge sufficient to know what expertise to look for in a collaboration will go a long way in helping make fulfilling contributions.

Sandhya Dwarkadas is the Walter N. Munster Professor and Chair of the Department of Computer Science at the University of Virginia. Until July 2022, she was the Albert Arendt Hopeman Professor of Engineering in the Computer Science Department at the University of Rochester, where was on the faculty since 1996 and served as chair from 2014-2020. She received her Bachelor's degree from the Indian Institute of Technology, Madras, India, and her M.S. and Ph.D. from Rice University. She is a fellow of the ACM and IEEE. She served as co-chair of the CRA-WP board from October 2019 – October 2022 and has been on the CRA-WP board since 2010. Her areas of research interest include parallel and distributed computing, computer architecture, and the interaction and interface between the compiler, runtime/operating system, and underlying architecture. She has made fundamental contributions to the design and implementation of shared memory both in hardware and in software, and to hardware and software energy- and resource-aware configurability.

RECENT ANNOUCEMENTS OF INTEREST TO THE VLSI COMMUNITY

(1) TSMC Expansion in Arizona to Target 3-nm Node

[https://www.eetimes.com/tsmc-expansion-in-arizona-to-target-3-nm-node/]

Taiwan Semiconductor Manufacturing Co. (TSMC) announced this week that it's building the factory shell for a possible second fab at its Arizona site. The world's top chipmaker has already committed to a $12 billion investment for its 5-nm fab in Arizona. For the second phase, it will move to a 3-nm node for delivery around 2025-2026.

(2) Nvidia's H100 Debuts in 'Henri,' Topping the Green500 List

[https://www.hpcwire.com/2022/11/14/nvidias-h100-debuts-in-henri-topping-the-green500-list/]

Nvidia's H100 GPU, the flagship of its Hopper architecture, has debuted on the Top500 and Green500 lists. The new GPU appears in the relatively small Lenovo-built Henri system, which also features Intel's Xeon "Ice Lake" CPUs. Perhaps most notably, though, Henri is now the world's most energy-efficient publicly ranked supercomputer, besting a bevy of Frontier-style systems that otherwise dominate the top ten of the Green500.

(3) Cerebras Builds 'Exascale' AI Supercomputer

[https://www.hpcwire.com/2022/11/14/cerebras-builds-exascale-ai-supercomputer/]

Cerebras is putting down stakes to be a player in AI cloud computing with a supercomputer called Andromeda, which achieves over an exaflops of "AI performance." The company called Andromeda one of the fastest AI systems in the U.S. The system strings together 16 CS-2 systems in a cluster, with a total of 13.5 million compute cores focused on AI. Each CS-2 system has a wafer-sized chip with 850,000 cores, which is considered the largest piece of silicon ever made. The Andromeda system has 96.8 terabits of internal bandwidth.

(4) Samsung Electronics Unveils Cutting-edge Memory Technology to Accelerate Next-generation AI

[https://semiconductor.samsung.com/newsroom/tech-blog/samsung-electronics-semiconductor-unveils-cutting-edge-memory-technology-to-accelerate-next-generation-ai/]

One of the important trends in the field of artificial intelligence (AI), Hyperscale AI is an AI which can replicate human thinking and decision-making by learning on its own. Doing so, it can accomplish incredible tasks, such as creating images based on human language prompts by comparing existing image analysis functions. However, computing for Hyperscale AI requires a massive amount of data that can cause bottlenecks if the DRAM capacity and bandwidth for data transference are not adequately supported for Hyperscale AI models. As a workable solution, Samsung Electronics is preemptively developing memory technology that can overcome this problem. Samsung approached these challenges by utilizing PIM (Processing-in-Memory) and PNM (Processing-near-Memory) technologies as solutions.

(5) IBM's Biggest Quantum Chip Yet Could Help Solve the Trickiest Math Problems

[https://www.popsci.com/technology/ibm-quantum-summit-osprey/]

At the IBM Summit this week, the company announced the checkpoints they've hit so far, including a newly completed 433-qubit processor called Osprey, and updated versions of their quantum software. Osprey, which is almost three times larger than a previous 127-qubit chip called Eagle, uses many of the same technologies and designs, like a hexagon lattice structure on the chip surface that holds all the qubits. But 400 qubits can be a lot to manage, so engineers are constantly experimenting with fabrication techniques or small changes in design to make the processors less noisy and more efficient.

(6) Single-Laser and Optical-Chip Pair Sets Data Transmission Record

[https://www.photonics.com/Articles/Single-Laser_and_Optical-Chip_Pair_Sets_Data/a68478]

An international team of researchers has reportedly set a data transmission record using just a single laser and optical chip to transmit more than 1 Pbit/s. The researchers, from the Technical University of Denmark (DTU) and the Chalmers University of Technology in Sweden, transmitted data at 1.8 Pbit/s.

TCVLSI Sponsored Conferences for 2022

Financially sponsored/co-sponsored conferences

- ARITH, IEEE Symposium on Computer Arithmetic
  - ARITH 2022: http://arith2022.arithsymposium.org/ Virtual conference dates : Sept 12-14 2022
- 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