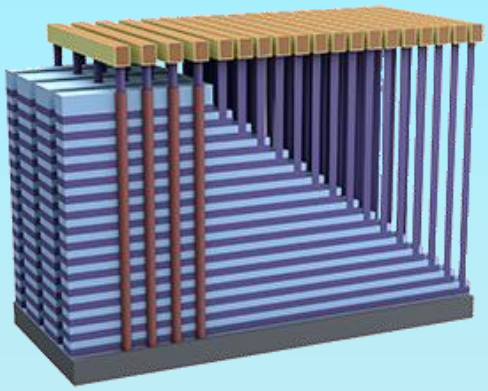
IEEE VLSI Circuits & Systems Letter

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

Volume 7 Issue 4 Nov 2021



Editor-in-Chief:
Mondira Pant

IEEE
IEEE computer society
The IEEE Computer Society Technical Committee on VLSI

IEEE VLSI Circuits and Systems Letter

Volume 7, Issue 4, Nov 2021

Editorial

Invited articles

- "The future of NAND flash – compute and non-volatile memory fusion" - *Sean Eilert (Micron Technology), Steffen Hellmold(Western Digital), Steve Kramer (Micron Technology), Victor Zhirnov (SRC)*

Conference spotlight

- ASYNC 2021 – Report by Program Co-Chairs -*Georgios D. Dimou (Niobium Microsystems, Inc.) and Milos Krstic (IHP - Leibniz Institut für innovative Mikroelektronik)*
- ASYNC 2021 Best Paper "Towards Explaining the Fault Sensitivity of Different QDI Pipeline Styles" - *Patrick Behal, Florian Huemer, Robert Najvirt, Andreas Steininger, Zaheer Tabassam*
- ASYNC-2021 Best Paper Nominee "Fluid: An Asynchronous High-level Synthesis Tool for Complex Program Structures" – *Rui Li, Lincoln Berkley, Yihang Yang, and Rajit Manohar*
- ASYNC-2021 Best Paper Nominee: "A 28nm Configurable Asynchronous SNN Accelerator with Energy-Efficient Learning" - *Jilin Zhang, Mingxuan Liang , Jinsong Wei, , Shaojun Wei , and Hong Chen*

Women in VLSI (WiV) series spotlight

- Interview with Dr. Alice Wang, Edge Platform Architect and VP of Hardware at Everactive, A startup that builds batteryless wireless sensors networks for Industrial IoT applications

Updates

- Recent relevant news highlights – *Ishan Thakkar*
- 2021 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 "***The future of NAND flash- Compute and non-volatile memory fusion***" by Sean Eilert (Micron Technology), Steffen Hellmold (Western Digital), Steve Kramer (Micron Technology), Victor Zhirnov (SRC), where the authors remind folks that in a world where data creation is increasing at an exponential rate, the future NAND will combine logic and memory on a single silicon to help balance cost scaling and performance.

The newsletter spotlights one of TCVLSI's sponsored symposiums in 2021, IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC) where researchers presented their latest findings in the area of asynchronous design. One-page teasers of the best paper awarded at ASYNC 2021 and two best paper nominees are showcased: "***Towards Explaining the Fault Sensitivity of Different QDI Pipeline Styles"; "Fluid: An Asynchronous High-level Synthesis Tool for Complex Program Structures"; "A 28nm Configurable Asynchronous SNN Accelerator with Energy-Efficient Learning"***

In our Women in VLSI (WiV) series, we share an inspiring interview with Dr. Alice Wang, Edge Platform Architect and VP of Hardware at Everactive, a startup that builds batteryless wireless sensors networks for Industrial IoT applications.

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 2021 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

The future of NAND flash – compute and non-volatile memory fusion!

Sean Eilert[1], Steffen Hellmold[2], Steve Kramer[1], and Victor V. Zhirnov[3]

[1]Micron Technology Inc.
[2]Western Digital Corp.
[3]Semiconductor Research Corp.

The Global DataSphere

Today, data is literally "eating the world." In fact, data is more than a megatrend—it can be viewed as a measure of the progress of humanity, cutting across all walks of life, public and private organizations, and every vertical of the economy.

The amount of data we create as a society is rising in an exponential manner. During 2020 alone, about 64 zettabytes of data was created and consumed[i]. This was further increased by the COVID-19 pandemic, which caused an upsurge in remote workers that now rely heavily on internet-based programs and video communication, as well as video recording and downloading. Over the next three years, the amount of data created will be more than the data created over the past 30 years combined. And, over the next five years, the world will create more than three times the data than it did in the previous five. These are, of course, enormous numbers as data is the fuel powering high-performance computing, social media platforms (and their reliance on data centers), edge computing (including autonomous driving), Internet of Things, and so many more.

Any attempt to capture the world's data production requires an understanding of the prefix "zetta-", which denotes a factor of 10^{21}, or a "billion trillion." According to the total bit inventory, presented in the Decadal Plan for Semiconductors[ii], mankind stored 1 zettabyte in 2010, somewhere between 10 and 100 zettabytes in 2020, and the world will store around 100,000 zettabytes in 2040. Of course, this is good news for the memory and storage industry. However, this huge number also represents a problematic challenge; global demand for data storage is growing so fast that, in the near future, today's technologies will not be sustainable due to the sheer mass of material resources needed to support the ongoing data explosion. The fact is, each bit has a weight—meaning, a weight basis in memory device silicon—and storing zillions of bits will necessarily zillions of grams of silicon and other materials.

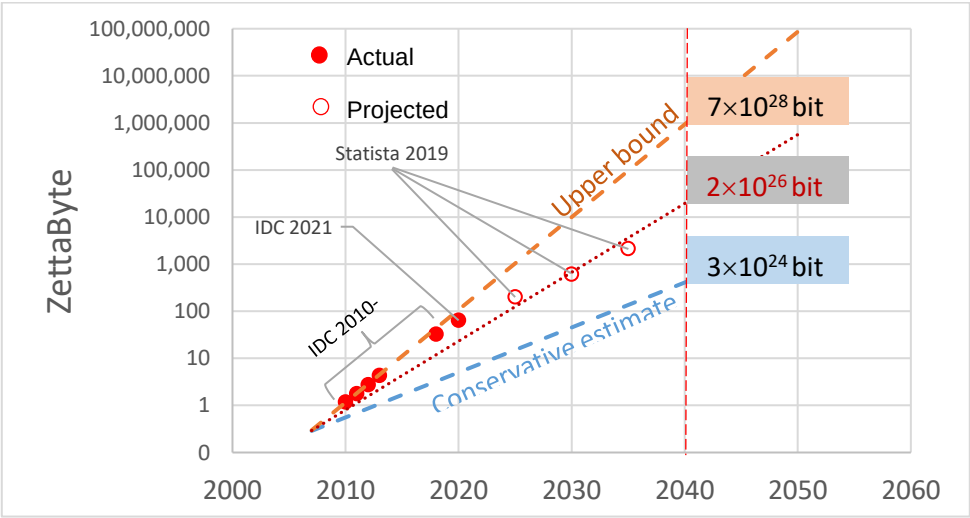


Figure 1. Global storage demand

How much does a bit weigh: 2D vs 3D NAND

As mentioned previously, it is important to remember that each bit of information is created in a materials system so there is a mass associated with it, and thus as with all man-made things, there is a material price or cost associated.

A bit of NAND Flash storage is among the smallest memory elements. In 2016 a group at Micron, Boise State, Harvard and SRC calculated the weight of one bit for the case of extremely scaled 2D NAND to be close to one picogram (10^{-12} g)[iii]. This is, of course, a tiny number – a typical mass of a bacterium cell! Now if we multiply this number by 2×10^{26} bit—the total data we expect to store in 2040 (Fig. 1) —it follows that the total mass of silicon wafers required would be more than 10^{10} kg. This, however, significantly exceeds the world's total available silicon supply (projected to be ~2.5×10^{7} kg of wafer-scale Si)! Obviously, we want to make bits as small as possible, but there are fundamental limits to scaling. For example, in electron-based devices, such as flash, quantum mechanics dictates a limit on material size—if the 'box' is too small, the electrons will too easily escape, and the information will be lost. This effect is known as quantum-mechanical tunneling and it forbids us from making a silicon bit less than about 15 nm in size (which corresponds to 10^{-12} g in weight). Therefore, we need to use other approaches to make get more bits. Indeed, the transition to 3D NAND manifested a revolution in storage, which allowed to significantly reduce the weight of bit. Let us do the numbers. NAND flash memory is organized into dense arrays (2D or 3D) of cells that are fabricated on a silicon wafer, and the mass of the wafer constitutes the largest component of a single cell mass. The initial wafer must be thick enough to avoid cracking during high-volume and speed manufacturing (for example a 300-mm wafer has a thickness h=0.775 mm). The weight of silicon needed for one flash memory bit, m_{bit}, is determined by the thickness of the wafer, h, the area of the flash cell, A_{cell}, and the density of silicon ρ_{Si} = 2.33 g/cm^3: $m_{bit} = \rho_{Si} \cdot A_{cell} \cdot h$. The area of one 2D NAND flash cell is A_{cell} = $4F^2$, where F is its smallest dimension. Based on the physics of operation of flash memory, F is 10-15 nm. Let, F = 15 nm, than yielding an A_{cell} = 9×10^{-12} cm^2. Assuming h = 0.775 mm for a 300 mm wafer, obtain m_{bit} = 1.63×10^{-12} g/bit. It should be noted that the thickness of an actual silicon die in a flash package is considerably smaller that the thickness of the initial wafer because the die thickness is reduced to 20–50 µm by back grinding and polishing. However, since the waste silicon from backgrind gets lost, the calculated weight is representative of the total amount of silicon required to build a 2D flash bit.

Two obvious concepts towards decreasing the weight of a bit include using multi-level cells that store more than one bit and organizing memory in three dimensions (3D). As an example, the 2021 estimate of a bit weight for the latest YMTC product (128 layers, TLC, 8.48 Gbit/mm^2) [iv] yields ~0.1 pg/bit, i.e. 10 times less than the limiting case of 2D NAND.

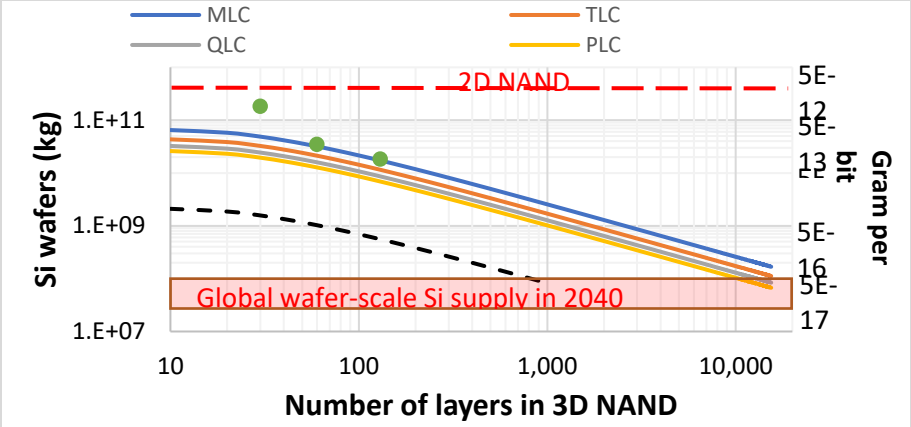


Figure 2. 3D NAND – What will be needed in 2040?
(MLC – 2 bits per cell, TLC – 3 bits per cell, QLC – 4 bits per cell, PLC – 5 bits per cell)

How far 3D NAND can go? - Challenges

Combining the 3D organization with multi-level storage allows to considerably decrease the weight of one bit. This is illustrated in Figure 2, where theoretical curves (idealized assumptions for 3D NAND scaling) are compared with several data points for NAND products[iv]. Clearly, very significant innovations in 3D NAND will be required in order to meet the projected 2040 global storage demand of about 2×10^{26} bit, given the anticipated Si supply limitations ($<10^8$ kg in 2040). For example, the needed number of layers will exceed 10,000. Also, as can be seen in Fig. 2., increasing the number of pits per cell above five (PCL) will gain only diminishing improvement. What 3D NAND will look like in 2040? Clearly, immediate steps of NAND scaling on the horizon are in the area of layer count as well as multi-level-cell NAND technology – unfortunately both of these levers providing only limited gains as the diminishing returns are already visible. A radical solution towards decreasing the weight of a NAND bit would be the recovery of silicon after backgrinding (the black dashed line in Fig. 2). This would allow to make the 2040 global storage goal with less than 1000 layers in 3D NAND and without a dramatic increase of Si wafer supply. Can the recovery of waste silicon from backgrind be economically feasible? A recent pilot project at TSMC shows promise[v]; of course a lot needs to be done to get to the recycling the backgrind waste back to the wafer-scale silicon. This will require both fundamental research investments and massive infrastructure developments.

More than just Storage – Compute with NAND?

Beyond density scaling, the future NAND will combine logic and memory on a single silicon with approaches such as CMOS under array (CUA) to help balance cost scaling and performance.

Memory is a cornerstone of modern computing systems, be it DRAM or extremely fast and expensive SRAM cache and scratchpad memory on, or tightly coupled to, processor silicon. Similar to the way DRAM is used for tiering and buffering data to avoid storage accesses, SRAM-based cache designs are used to avert costly accesses to main memory. SRAM collocated on processor ASICs has gone from 25% of total transistor count in 1990 to nearly 80% in 2020. In short, with the advent of multi-core processors and domain specific accelerators, memory subsystems are a dominant consideration in the design of modern computing systems.

Increasing working memory is a first step to driving compute advancements but ultimately the industry will need to intertwine NAND flash and compute given that NAND flash holds the greatest promise of scaling

in the hot domain and can deliver the best power efficiency if coupled with purpose build compute elements.

The cost of data movement can no longer be overlooked in modern systems and applications design. There is a hierarchy of memory and storage types ranging from processor registers and caches to main memory to storage and data movement through this hierarchy to the computing elements consumes far more energy than is required to complete the computation itself. Domain specific accelerators are being implemented with tightly coupled that offer reduced energy consumption, higher bandwidth and reduced latencies. Technologies such as in-memory-compute and byte-addressable storage are becoming increasingly attractive because they circumvent the need to perform costly data moves.

Emerging Vertical market failure risks: Call for action

An important final remark: While the memory and storage industries are experiencing unprecedented growth, the risks of vertical market failure have recently emerged. This is due in large part to increasing technical asset specificity where there are few buyers outside of hyperscale customers. Moreover, specialization often creates niche markets that are sub-scale in size, making significant R&D investments difficult to justify. This scenario creates a dilemma for manufacturers as they have historically carried both the burden of product research and development, as well as the subsequent commercial risk. The prospect of vertical market failure can be mitigated by private sector market participants through risk-share agreements between customers and suppliers, as well as increased vertical integration. Moreover, given the strategic nature of data management and storage, it would also be beneficial for the government to lend public-sector market leverage to help ensure that memory and storage innovation is supported to an extent necessary to spur needed innovation, and ensure America has a hand in the market.

The scaling of NAND flash in production as well as R&D development is facing increasing costs at multi-billion dollar scale which may represent increasingly difficult for suppliers. Given the government support for the local semiconductor industry in general as well as the local NAND industry in other regions there is a growing risk that the US may lose its leading edge in NAND manufacturing as well as technology innovation. Without doubt there is a significant NAND scaling opportunity ahead driving by major disruptions which represents an opportunity for those investing as well as a threat for the incumbents in case they are out-invested. Thus, it is vital that the US governments takes an active role in supporting the growth of the NAND manufacturing infrastructure as well as advanced technology developments to continue to secure a leadership position. Paraphrasing a famous quote, “who owns the NAND innovation and the ability to turn data stored in it into information, owns the world”.

References

[1] Data Creation and Replication Will Grow at a Faster Rate than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts, https://www.idc.com/getdoc.jsp?containerId=prUS47560321 (published 24 Mar 2021)

[2] SIA/SRC Decadal Plan for Semiconductors (SRC 2021) https://www.src.org/about/decadal-plan/

[3] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, W. L. Hughes, “Nucleic acid memory”, NATURE MATERIALS 15 (2016) 366-370

[4] Trends of 3D NAND 1GB Area, 2021 TechInsights Inc., https://www.techinsights.com/blog/memory/ymtc-128l-3d-xtacking-20-tlc-nand

[5] TSMC Pioneers Physical Regeneration Technique for Backgrinding Wastewater https://esg.tsmc.com/csr/en/update/greenManufacturing/caseStudy/44/index.html

Sean Eilert is a Micron Fellow in the Technology Pathfinding Group performing research at the intersection of emerging memories and the systems that will utilize them. Sean graduated with a Degree in Electrical Engineering from the University of Kansas and across his 30-year career he has worked for leading semiconductor companies such as Intel, Numonyx and Micron. Sean has held numerous engineering development roles ranging from test development, reliability, media management, design, architecture, system architecture, system design and system characterization. Sean's interests lie in memory architectures, system architectures, and the interactions between them with special interest in media management and in-memory compute. An innovator by nature, Sean holds over 40 memory, compute-in-memory and system-related patents. Outside of his professional career, Sean enjoys a variety of activities ranging from making traditional Japanese wood-fired ceramics to yoga to outdoor rock climbing.

Steffen Hellmold is Vice President of Strategic Initiatives at Western Digital. He is responsible for identifying & leading incubated disruptive innovation projects in support of storage adjacent growth opportunities. His current area of focus is Archive Storage with an emphasis on driving tape and DNA Data Storage innovation to address the growing demand for cloud-scale accessible and deep archive solutions providing improved density and longevity as well as lower product acquisition and operating costs.

Before joining Western Digital in 2013, Hellmold was Vice President of Marketing at Everspin Technologies. Prior positions include executive management at SandForce, Seagate Technology, Lexar Media (a subsidiary of Micron), Samsung Semiconductor, Fujitsu and SMART Modular Technologies. Hellmold has been deeply engaged in various industry trade associations as well as industry standards organizations such as JEDEC, IEEE, USB-IF, CFA, SDA, and MMCA. He co-founded the USB Flash Drive Alliance and served as their president from 2003 to 2007. Hellmold holds an Economic Electrical Engineering degree (EEE) from the Technical University of Darmstadt, Germany

Steve Kramer is a Principal Engineer and serves as a coordinator for external research in the Technology Development Group, helping to identify and vet new semiconductor memory technology.Steve started his career at Micron doing chemical mechanical planarization (CMP) where he helped standardize in-situ measurement techniques for enhancing process control. He followed one of his side-projects, related to supercritical fluid deposition, into technology pathfinding and university project liaison work. This knack for technology dabbling has also led him to garner more than 50 patents and to publish a research paper from time to time. Initially earning a degree in Chemistry from the College of Idaho, Steve subsequently achieved an advanced degree in Materials Science from the University of California, Los Angeles, researching chemical routes to ceramics and flexible aerogels. Steve has also had the chance to serve as a visiting researcher at Nippon Sheet Glass, Japan, studying non-wettable coatings, and as a Research Scholar at Stanford University, investigating the potential of magnetic tunnel junctions for memory applications. In his free time, Steve enjoys a plethora of outdoor enterprises, including snow skiing, mountain biking, rock climbing, and elk and mushroom hunting

Victor Zhirnov is Chief Scientist at the Semiconductor Research Corporation. He is responsible for envisioning new long-term research directions in semiconductor information and communication technologies for industry and academia. His semiconductor experience spans over 30 years in the areas of materials, processes, devices physics and fundamental limits. Victor served as the Chair for the Emerging Research Device (ERD) Working Group for the International Technology Roadmap for Semiconductors (ITRS) as well as for the Semiconductor Synthetic Biology and Bioelectronic Medicine Roadmaps. Victor received the M.S. in applied physics from the Ural Polytechnic Institute, Ekaterinburg, Russia, and the Ph.D. in solid state electronics and microelectronics from the Institute of Physics and Technology, Moscow, in 1989 and 1992, respectively. He has authored and co-authored over 150 technical papers and contributions to books.

Conference Report - ASYNC 2021

Georgios D. Dimou (Niobium Microsystems, Inc.) and Milos Krstic (IHP - Leibniz Institut für innovative Mikroelektronik), Program Co-Chairs

The International Symposium on Asynchronous Circuits and Systems (ASYNC) is the premier forum for researchers to present their latest findings in the area of asynchronous design. The conference addresses topics that range from theory of asynchronous system operation, to fundamental circuits and tools that enable asynchronous design and all the way to applications of such principles to build complex systems. The 27th symposium was hosted by Galois, Inc. and held remotely September 7-10, 2021.

The conference received 19 regular paper submissions that have entered the review process. The submissions were reviewed by the 39-member technical committee. Ultimately the committee accepted 9 papers for publication. Moreover, the program was extended with 3 short fresh ideas papers, which have presentations at the conference, but are not included in the proceedings. The conference also had an invited session for best paper nominees from ASYNC 2020, which was cancelled last year due to the COVID-19 related circumstances.

ASYNC 2021 paper presentations were split over 3 regular sessions, each one of which was led by keynote presentations, as well as a fourth session that was dedicated to the best paper awards for both 2020 and 2021. The keynote presentations were split between industry and academia with world-class presenters on both sides. Specifically, Edith Beigne (Facebook) presented on the future of AR/VR platforms, Prof. Kwabena Boahen (Stanford University) on neuromorphic computation and his vision for a 3D silicon brain and Dr. Ran Ginosar (Technion & Ramon Space) on space applications.

The ASYNC 2021 conference was attended by 50 online participants from 13 countries around the world, both from academia and industry. The conference also received support from sponsors NVidia and Galois, as well as from the IEEE Computer Society/TCVLSI.

Towards Explaining the Fault Sensitivity of Different QDI Pipeline Styles

Patrick Behal, Florian Huemer, Robert Najvirt, Andreas Steininger, Zaheer Tabassam

Institute for Computer Engineering, TU Wien, Vienna, Austria

Asynchronous circuits, specifically those using a quasi delay-insensitive (QDI) implementation are known for their high resilience against timing uncertainties. However, their event based operation principle impedes their temporal masking capability, making them more susceptible to fault-induced transitions caused by single event transients. While synchronous circuits obtain high resilience through temporal masking that is established through the sampling of data by flip flops, asynchronous circuits, by design must be flexible about the phases of data validity leaving a larger attack surface for faults. Consequently, previous work has proposed to narrow down the windows in which data changes are accepted, in order to improve the temporal masking in QDI designs. Unfortunately, it is hard to combine the existing insights and results into a global picture, as they all have been derived for specific circuits and pipeline types, under specific experimental conditions (or by theoretical analyses) and with specific targets in mind.

Our vision is to elaborate such a global picture through a large experimental study (complemented by theory) that allows an apples-to-apples comparison of different pipeline styles and fault-tolerance enhancements. Since the masking effects in asynchronous design seem to depend on many operational parameters like the pipeline fill level, or the data being processed, another target is the identification of such factors along with a modeling of their specific influence. On the foundation of this understanding, we can then identify the main vulnerabilities, the most efficient existing enhancement approaches, and finally elaborate further improvements. In this paper we report about some first important steps in this direction. We present an experimental environment that allows the convenient generation of target circuit descriptions, as well as the fully automated conduction of large gate-level simulation experiments with millions of fault injections, while still providing the ability to precisely reproduce each single fault injection for closer inspection of interesting cases. Using this tool we perform a detailed comparison of fault effects seen in different QDI pipeline styles (variants of the weak condition half buffer (WCHB) and Mousetrap-style D-latch half buffer), for a multiplier with varying bit width implemented in delay-insensitive minterm synthesis (DIMS) with randomized delays. Beyond different levels of pipelining we also consider an iterative circuit topology, and we vary the pipeline fill level in several steps from token limited to bubble limited.

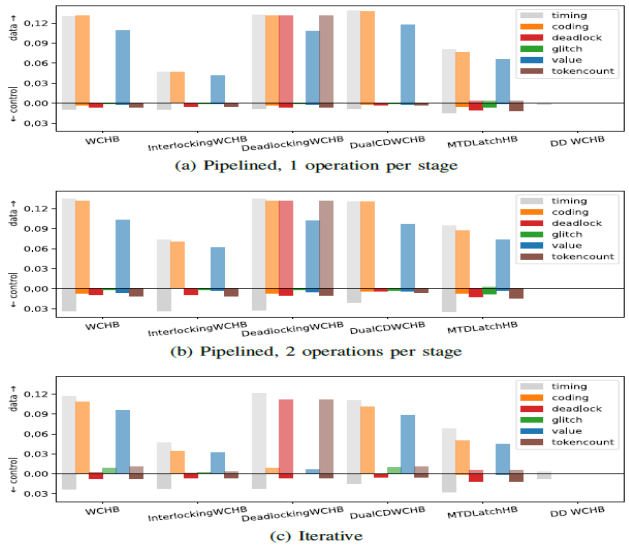


Figure 1: Number of observed effects relative to the total number of injections for 8 bit multiplier designs

Figure 1 shows an example of our results, where, e.g. good robustness of the interlocking WCHB can be observed, low sensitivity of control signals relative to data signals, as well as interesting differences between linear pipeline and iterative implementation. More generally, our results allow to study the relative sensitivity of circuits resulting from different design choices.

Reference:

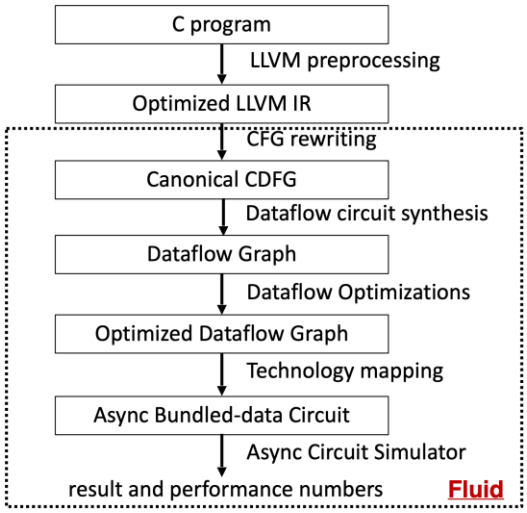
[1] Patrick Behal, Florian Huemer, Robert Najvirt, Andreas Steininger, Zaheer Tabassam, "Towards Explianing the Fault Sensitivity of Different QDI Pipline Styles", in 2021 IEEE International symposium on Asynchronous Circuits and Systems (ASYNC2021), Sep 7-10, 2021, USA -Best paper award

Fluid: An Asynchronous High-level Synthesis Tool for Complex Program Structures

Rui Li, Lincoln Berkley, Yihang Yang, and Rajit Manohar

Computer System Lab, Yale University, New Haven, CT 06520, USA

There have been significant efforts in high-level synthesis (HLS) to translate behavioral descriptions of algorithms into inherently pipelined dataflow circuits. The most complex aspect of generating dataflow circuits is managing conditional execution and conditional token generation. However, prior works either mostly avoid conditional tokens, or only support conditional tokens for simple control structures. As a result, these tools have difficulty generating fully-pipelined dataflow circuits, or only support a subset of programs. Fluid is a HLS tool that maps C programs with complex control structures into fully-pipelined asynchronous dataflow circuits. It also applies several dataflow optimizations on the generated circuits to improve the performance.



Fluid is built on top of the LLVM compiler framework. LLVM maps C programs into optimized LLVM Intermediate Representation (IR), from which Fluid constructs a control-data flow graph (CDFG) that captures the control and data dependencies of the C programs. Then, Fluid converts it into canonical form, which consists of sequential blocks, properly nested conditional blocks and canonical loop blocks (with single enter/exit point). After that, the canonical CDFG is synthesized to a dataflow circuit (represented as a dataflow graph) which is composed of concurrent dataflow elements[1]. Next, Fluid applies dataflow optimizations on the graph to try to cluster dataflow elements together for more logic optimization opportunities. Lastly, Fluid maps the optimized dataflow graph into asynchronous bundled-data circuits. It uses commercial tools to design the combinational function units and augments them with asynchronous control logic. We also built a discrete-event asynchronous circuit simulator to verify the circuit results and measure the performance.

One key challenge solved by Fluid is handling CDFGs with irregular control structures, which arise in real-world applications with complex control statements (e.g, **return** inside *if* statement, **break**/**return** inside *loop* statement, etc.), as well as compiler optimizations. Note that the canonical CDFG requires that each if/loop statement should have a single entry and a single exit. To handle this, a programmer could introduce a new *flag* variable and rewrite the program with the *flag* variable to get rid of the complex control statements. Inspired by this, Fluid proposes control-flow graph rewriting algorithms that directly convert the non-canonical IR into canonical form. We also proved the convergence of the CFG rewriting algorithm.

We compare Fluid with the open-source Legup HLS tool as well as two commercial HLS tools. We measure the delay, area, energy, leakage power and throughput of the generated circuits. No HLS tool is consistently the best in all metrics across all benchmarks, and the key is to achieve a good tradeoff among these metrics. Compared with academic tools, Fluid results in 8.33X reduction in energy, 2.5X increase in throughput, 1.64X improvement in delay at the cost of 1.19X (1.95X) increase in area. Compared with commercial tools, Fluid results in 3.58X reduction in energy, 1.34X increase in throughput at the cost of 1.15X increase in delay and 1.95X increase in area.

Reference:

[1] John Teifel and Rajit Manohar," ***Static Tokens: Using Dataflow to Automate Concurrent Pipeline Synthesis***", 10th International Symposium on Advanced Research in Asynchronous Circuits and Systems – Best paper award nominee

A 28nm Configurable Asynchronous SNN Accelerator with Energy-Efficient Learning

Jilin Zhang*, Mingxuan Liang*, Jinsong Wei†, Shaojun Wei*, and Hong Chen*

*School of Integrated Circuits, Tsinghua University

†School of Microelectronics, University of Science and Technology of China

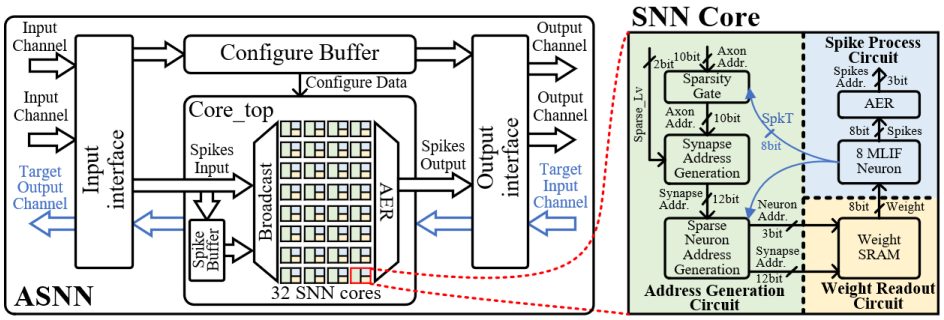


Fig. 1. Overall architecture and core structure of our SNN accelerator.

With the development of spike neural network (SNN), more and more SNN accelerator has been deployed in smart devices. However, SNN accelerators face three main challenges: 1) Energy efficiency, high energy efficiency means long battery life, which is critical for energy-constrained device. 2) On-chip learning ability, which enables the chip to learn from the environment for different applications. 3) Configurability, SNN accelerator needs to be configurable to work with different topology to meet the requirement of various applications.

Our accelerator is proposed with three key contributions: 1) SNN is implemented with asynchronous circuits for spikes processing. The event driven nature of asynchronous circuits allows the processor consumes energy only when and where needed. 2) A sparse target propagation (S-TP) algorithm is proposed, which achieves up to 99.11% hand gesture recognition accuracy in complex background and lighting conditions. 3) A skip-training mechanism is put forward to skip 90% on-chip training process to reduce the training energy overhead.

As shown in Fig. 1[1], our asynchronous SNN accelerator consists of four parts: input interface, output interface, configure module, and Core_top module. The input and output interfaces are parallel-to-serial and serial-to-parallel convertors respectively. In Configure Buffer module, the configure information (such as train enable signal and SNN core enable signal) is stored and transmitted in Core_top module. Core_top module is the main part of the SNN accelerator, which consists of 32 cores, each core has 8 neurons and 4096 synapses. Spike Buffer in Core_top module is used to store the spike traces and release the spikes to SNN cores during training.

The accelerator is tapped out in 28-nm CMOS process. The accelerator reaches 95.7% accuracy on NMNIST dataset and 99.11% accuracy on non-stanard dynampic hand getsure dataset with on-chip learning. When accelerator runs on real-time, training power consumption (160.5μW) is only 0.56% more than the inference power consumption (159.6μW).

Reference:

[1] Jilin Zhang, Mingxuan Liang, Jinsong Wei, Shaojun Wei, and Hong Chen, "**A 28nm Configurable Asynchronous SNN Accelerator with Energy-Efficient Learning**", in 2021 IEEE International symposium on Asynchronous Circuits and Systems (ASYNC2021), Sep 7-10, 2021, USA -Best paper award nominee

WOMEN-IN-VLSI (WiV) SERIES: Dr. Alice Wang

Dr. Wang is currently the Edge Platform Architect and VP of Hardware at Everactive, a startup that builds batteryless wireless sensors networks for Industrial IoT applications that operate off of renewable energy.

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

Q1. What VLSI area have you focused on?

My PhD at MIT was in ultra-low power circuit design back when low-power was still in its infancy and was the seminal paper in Subthreshold digital design. Since then I've worked on developing Low-power High-performance CPU's for Smartphones at companies like Texas Instruments and MediaTek. It's is a full circle to be able to join Everactive which was started by my PhD buddies. We care about every nano-Watt of power because our products are operating off of harvested energy.

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

Given the urgency of climate change and wanting to take care of the environment we really do need to take a closer look at how we are using energy. I know that I take for granted that the lights will go on and at times careless leave them on without thinking about the impact on our world. There is more and more demand for computing and information. One less battery or fewer fossil fuels burned can make a big difference for world and future generations.

Q3: What is your typical day like in your current role?

In addition to developing the architecture for our platform offering, I also enjoy picking up new technology projects for the company and seeing them to completion. Every day I look around and pick activities that bring me joy and help me to learn new things. Some days I am looking at our sensors and doing some data analysis to see if they are surviving in the harsh environments or able to wake up after not having power all night. Other days I am mentoring young managers and help train them in soft skills. I found that my technical education did not prepare me for management and people leadership, so I want to make sure our young leaders are ready for the challenge.

Q4: What motivated you into this field?

I kind of stumbled into electrical engineering. In high school, I was convinced that I would be a physicist since I had the highest grades in AP Physics. At MIT, a few semesters in, I realized my brain didn't think like the other physics students did. So I took a couple of EE classes and switched after I saw how fun it was to code and play with circuits to build things. After my M. Eng, I decided to stay in school and "play" for a while longer while getting my PhD. I chose VLSI and microelectronics because I thought if I was dealing with 10's and 100's of Watts, I might accidentally kill myself so for safety sake I stuck with micro- and nano-Watts low power design.

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

I have a "bucket list" of things I want to work on or do in my lifetime. I started this list about 15 years ago and it's pretty amazing to see some of the things come true such as live and work outside of the US (I was an expat in Taiwan for 4 years) and work at a startup (now at Everactive). One thing on my list that I am very hopeful about is to work for the first female President of the United States. I am pretty sure a female US President will happen in my lifetime, just not sure if I'll get a chance to work for her directly.

The other thing high up on my list is to cure cancer. For years, I've always joked to myself that we need to find a cure for cancer because I'll probably get it since I don't eat organic, don't enjoy exercising, etc.

Well the joke is on me when last year right before the global pandemic I was diagnosed with early stage breast cancer. I went through chemo and surgery and now am cancer free and a survivor. One thing I realized through the experience is that there are lots of cures for cancer and it's not a death sentence. Just that the cures are very tough to endure and I would never wish to see anyone else have to go through it. So, I am amending my new goal to eradicating cancer!

Q6: Many students of color and women worry that the VLSI STEM field won't welcome them. When you look at the landscape for scientists and engineers, what do you see right now? Are there signs of progress?

When I entered the field of VLSI STEM, during my studies I was able to take advantage of various scholarships and fellowships that promoted women. I felt I had a lot of female friends who provided support. As I rose up the corporate ladder it did get more and more lonely with fewer women in meeting rooms and fewer female colleagues at work. What has been encouraging in the last few years is that there is a bright light shining on the problem and a lot more efforts to increase diversity, inclusion and equity at all levels. Where it used to be the few underrepresented employees trying to make minute changes, now there are a lot more male allies lending their support.

I am currently the chair of the Women in Circuits for the Solid-State Circuits Society. We started in 2017 with the mandate from the SSCS President Jan Van der Speigel and were even given a budget. Even though we all knew each other for years it wasn't until we were given a charter and a gentle push did we start to work together as a team to recruit, retain and advance women. Previously, each woman was trying their best as the "only" in their organizations. Now we have a community of women who supports each other.

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

Until the pandemic, I enjoyed having a completely unbalanced life. I really love working and relished having my schedule crammed packed with work, travel, volunteering with IEEE, serving at church, meeting with friends and my Netflix queue. It was right before the pandemic happened when I received my cancer diagnosis and for the first time I had to put health before everything else. Because I knew immediately that I would need emotional support, I requested to work remotely for the duration of my treatment and moved to be close to family. I went from a booked schedule to part time and didn't push myself to log in unless I was feeling 100%. Since then I have tasted what it's like to have a balanced life and make sure to incorporate more joy practices into my life. I take a look at my schedule every week and remove all unnecessary meetings. I deleted work emails and Slack from my phone (trust me this was revolutionary). Then I had room for daily walks, personal training at the gym, experiments with the air fryer and virtual meetings with friends.

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

Spend time knowing yourself, what you want and who you are. Trust yourself and live every day like it's your last. Rinse and repeat!

Alice Wang received her Bachelors, Masters, and Ph.D. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, in 1997, 1998, and 2004, respectively. She wrote the paper "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology" with Professor Anantha Chandrakasan which inspired a new research field in ultra-low power technology. After her PhD, she spent 8 years at Texas Instruments developing low-power circuit and system technology for mobile, application processors and radios. At MediaTek, Alice was an Assistant General Manager in High-Performance Processor Technology working on SoC's and CPU's for consumer electronics including Smartphones, Tablets and Smart TV's and managing the Foundation IP teams (Standard cell library, Memory and I/O). Her work on low-power technology has been showcased in 30+ IEEE publications and she has co-authored two books. She is a Senior Member of IEEE. Currently, Alice is the Platform Architect and the Vice President of Hardware at Everactive working on self-powered sensing for Industrial IoT applications. She was elected to the Advisory Committee for the Solid State Circuits Committee (2017-2019) and is a Founding member of the Women in Circuits committee.

(1) Intel Unveils Loihi 2, Its Second-Generation Neuromorphic Chip
[https://www.hpcwire.com/2021/09/30/intel-unveils-loihi-2-its-second-generation-neuromorphic-chip/]

Four years after the introduction of Loihi, Intel's first neuromorphic chip, the company is introducing its successor. The second-generation chip will provide faster processing, higher resource density and greater energy efficiency. Intel is also introducing Lava, a software framework for neuromorphic computing.

(2) AMD's Multi-Chip MI200 GPU Readies for a Major Global Debut
[https://www.hpcwire.com/2021/10/21/killer-instinct-amds-multi-chip-mi200-gpu-readies-for-a-major-global-debut/]

AMD's next-generation supercomputer GPU is on its way – and by all appearances, it's about to make a name for itself. The new GPU is based on AMD's CDNA2 architecture and uses a mezzanine form factor. The AMD Radeon Instinct MI200 GPU (a successor to the MI100) will, over the next year, begin to power three massive systems on three continents: the United States' exascale Frontier system; the European Union's pre-exascale LUMI system; and Australia's petascale Setonix system.

(3) TSMC Details the Benefits of Its N3 Node
[https://www.eetimes.com/1383768-2/]

The N3 node, which will provide more of a technological leap than N4, is planned to go into volume production in the second half of 2022. N3 will indeed offer customers the kind of performance improvements they might hope for from a major node jump. Going to N3, customers would get a 10% speed boost at 26% less power.

(4) HBM3: Big Impact on Chip Design; new levels of system performance bring new tradeoffs
[https://semiengineering.com/hbm3s-impact-on-chip-design/]

An insatiable demand for bandwidth in everything from high-performance computing to AI training, gaming, and automotive applications is fueling the development of the next generation of high-bandwidth memory. HBM3 will bring a 2× bump in bandwidth and capacity per stack, as well as some other benefits. The memory technology is becoming significantly faster and wider. In some cases, it is even being used for L4 cache.

(5) UCLA Researchers Report Largest Chiplet Design and Early Prototyping
[https://www.hpcwire.com/2021/10/12/ucla-researchers-report-largest-chiplet-design-and-early-prototyping/]

The team of researchers has designed and is now prototyping a "2048-chiplet, 14336-core waferscale processor". This is the largest chiplet assembly-based system ever attempted. In terms of active area, the prototype system is about 10x larger than a single chiplet-based system from Nvidia/AMD etc., and about 100x larger than the 64-chiplet Simba (research) system from Nvidia.

(6) NeuroBlade Raises Funds to Launch its Compute-in-Memory Chip
[https://www.eetimes.com/neuroblade-raises-funds-to-launch-its-compute-in-memory-chip/]

NeuroBlade, the compute-in-memory startup, has secured $83 million to help market its data analytics accelerator based on its XRAM computational memory chip. The Israel-based startup has developed a data analytics architecture that eliminates major data movement bottlenecks by integrating data processing functions in-memory.

(7) Toshiba shrinks quantum key distribution technology to a semiconductor chip
[https://www.toshiba.eu/pages/eu/Cambridge-Research-Laboratory/toshiba-shrinks-quantum-key-distribution-technology-to-a-semiconductor-chip?platform=hootsuite]

Toshiba Europe Ltd today announced it has developed the world's first chip-based quantum key distribution (QKD) system. This advance will enable the mass manufacture of quantum security technology, bringing its application to a much wider range of scenarios including to Internet of Things (IoT) solutions.

TCVLSI Sponsored Conferences for 2021

Financially sponsored/co-sponsored conferences

- ARITH, IEEE Symposium on Computer Arithmetic
  - ARITH 2021: http://arith2021.arithsymposium.org/ Virtual conference dates : June14-16 2021
- ASAP, IEEE International Conference on Application-specific Systems, Architectures and Processors
  - ASAP 2021: http://2021.asapconference.org/ Virtual conference dates: July 10-12 2021
- ASYNC, IEEE International Symposium on Asynchronous Circuits and Systems
  - ASYNC 2021: https://asyncsymposium.org/async2021/ Virtual conference dates: Sept 7-10 2021
- iSES, (formerly IEEE-iNIS) IEEE International Smart Electronic Systems
  - IEEE iSES 2020: https://ieee-ises.org/2020/ises-cfp/ December 14-16, 2020, Chennai, India
  - IEEE iSES 2021 Dec 20-22 2021
- ISVLSI, IEEE Computer Society Symposium on VLSI
  - ISVLSI 2021: http://www.eng.ucy.ac.cy/theocharides/isvlsi21/ Virtual conference dates: July 7-9 2021
- IWLS, IEEE International Workshop on Logic & Synthesis – collocated with DAC
  - IWLS 2021: Virtual conference dates: June 19th-21st, 2021
- SLIP, ACM/IEEE System Level Interconnect Prediction
  - SLIP 2021: https://dl.acm.org/conference/slip/proceedings Date TBD

Technically Co-Sponsored Conferences for 2021

- VLSID, International Conference on VLSI Design
  - VLSID 2021: https://embeddedandvlsidesignconference.org/ Virtual conference dates: Feb 20 -24 2021

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