Supercomputers and European Sovereignty

Dr. Mateo Valero Cortés
Barcelona Supercomputing Center (Director)
Spain

Abstract

Over that last 3 decades, we have witnessed a transition from closed software ecosystems being the foundation for HPC, enterprise, and business to open source software ecosystems based on Linux: from Arduino in the IoT space, to Android in the mobile space to Linux in HPC and cloud-based systems with various Open Source Software projects built on top. However, when examining hardware, current commercial off the shelf solutions are closed hardware ecosystems that only enable integration at the peripheral (PCIe) level. The combination of current technology trends, the slowing of Moore's Law, and cost prohibitive silicon manufacturing inhibit significant power-performance gains by relying on traditional closed ecosystems, especially in HPC, technology pushed to the extreme. This new regime forces systems to be much more specialized to achieve the power-performance profiles required for a supercomputer. In the past, HPC has led the way forward, defining the bleeding edge of technology. HPC can do this again with open hardware, as it has done in software with adopting Linux and open source in general. This is not only a technology imperative, but one born out of current geopolitics. Digital Technology (the generation and processing of data) is the basis for global commerce, scientific discovery, and ubiquitous in modern life. Thus, creation of digital technology in the form of processors, accelerators and the related digital infrastructure guarantees access to these building blocks of the digital economy regardless of the geopolitical environment. Given this technology and geopolitical backdrop, we describe how Europe can exploit its resources targeting research and development for technological independence.

In this today's technology environment, some of the rules have changed. This has produced a shift from abundant transistors to efficient use of transistors. Thus, to truly meet the power and performance requirements, we must specialize the hardware. At the same time, the software stack is evolving, becoming more abstract, enabling higher programmer productivity, but sacrificing hardware efficiency. Thus, application owners will need to co-design the full stack, all layers of hardware and software, in order to meet their performance and power (e.g., FLOPs/W) targets. This level of integration is not possible in a closed or even partially open ecosystem. The platform must be open to enable this tight integration. We see this openness today in the Linux OS, toolchain, runtimes, frameworks, and libraries, up to the application layer. This enables rapid development and extension of software systems. However, an open hardware infrastructure was lacking, making specialization nearly impossible, especially in a research context. Openness is required to tailor your hardware platform to the applications, thereby achieving the desired performance in the power constrained environment. There have been a couple of open source hardware platforms in the past, but Moore's Law inhibited their adoption for many reasons: general purpose processor improvements, time to market, cost, software development, etc. Furthermore, unlike Linux, previous open source hardware was entangled in the companies that created them. Mirroring the same model as Linux, RISC-V has followed a similar development path and has enjoyed significant industrial and academic adoption. Like Linux before it, the RISC-V ecosystem is in the nascent period where it can become the de facto open hardware platform of the future. The RISC-V ecosystem has the same opportunity in hardware that Linux created as a foundation for open source software. This enables the co-design of the RISC-V hardware and the entire software stack, creating a better overall solution than the closed hardware approach that is done today. RISC-V enables everyone to build what they want and need vs. buy something that partially meets their

requirements. As European HPC recognized in the past with Linux, Europe has the opportunity to lead the charge, creating a full stack solution for everything from supercomputers to IoT devices, all based on an open ISA, providing interoperability and a freedom to create, build, and deploy superior technology based on European IP.

In this talk, first, we will provide background on HPC computing and the research we have conducted to shape the current state of the art in HPC. Using RISC-V as an instrument, we provide a vision for the future and a collection of current research and innovation projects, infrastructure, and the community that are building the foundation for the future. This is a new opportunity for Europe to lead the way to an HPC Future that is Wide Open!

Biography

Mateo Valero, http://www.bsc.es/cv-mateo/. Director of the Barcelona Supercomputing Center. His research focuses on high performance architectures. He has published approximately 700 papers, has served in the organization of more than 300 International Conferences and he has given more than 600 invited talks. Prof. Valero has been honored with several awards, among them the 3 most relevant awards in Computer Architecture field: The Eckert-Mauchly Award 2007 by the IEEE and ACM, the Seymour Cray Award 2015 by IEEE and the Charles Babbage 2017 by IEEE. Among others: The Harry Goode Award 2009 by IEEE, The Distinguish Service Award by ACM and the Spanish National awards “Julio Rey Pastor” and “Leonardo Torres Quevedo”. "Hall of the Fame" member of the ICT European Program (selected as one of the 25 most influents European researchers in IT during the period 1983-2008, Lyon, November 2008). In 2020 he has been awarded for his exceptional leadership in HPC by “HPCWire Reader’s Choice Awards” for “being an HPC pioneer since 1990 and the driving force behind the renaissance of European HPC independence”. Honored with “Condecoración de la Orden Mexicana del Águila Azteca” 2018, highest recognition granted by the Mexican Government. He is Honorary Doctorate by 10 Universities. He is member of 9 academies. He is a fellow of IEEE and ACM and he is also Intel Distinguished Research Fellow.

In 1998 he won a “Favourite Son” Award of his home town, Alfamén (Zaragoza) and in 2006, his native town of Alfamén named their Public College after him.

AI-Assisted Yield Learning

Yu Huang
Huawei Semiconductor Scientist & HiSilicon EDA Chief Architect
HiSilicon Technologies Co., Ltd. – Shenzhen, Guangdong, P.R. China

Abstract

Root Cause Analysis (RCA) and Layout Pattern Analysis (LPA) are critical technologies for Diagnosis Driven Yield Learning in designing and manufacturing integrated circuits. Recent advancements of AI technologies can help improving yield learning accuracy and transferring the yield learning experiences from old designs to new designs or from old technologies to the new ones.

In this talk, we share our experiences in this research area and discuss the following techniques:

(1) A neural-network-based framework for RCA. The framework has a self-adaptive module that is able to adapt the inference module to new designs and new technologies based on a few new samples.

(2) An encoder network framework for LPA. It applies Contrastive Learning to extract representations of layout snippets that are invariant to trivial transformations such as shift, rotation, and mirroring. The layout snippets are then clustered to form layout patterns. The causal relationship between any potential layout patterns and the systematic defects is identified by the Causal Representation Learning.

(3) An unsupervised learning framework by using a Deep Latent Variable model consisting of a probabilistic encoder and a regularization decoder. The encoder transforms the features from diagnosis reports to latent variables characterizing the root cause distribution. The regularization decoder uses a Graph Attention Network to represent the mapping from the true root causes to suspicious root causes reported.

Biography

Dr. Yu Huang is Semiconductor Scientist at Huawei Technologies Co., and EDA Chief Architect of HiSilicon. Before Joining HiSilicon, he was Sr. Key Expert of Mentor Graphics (now Siemens EDA). His research interests include VLSI SoC testing, ATPG, compression, diagnosis, yield analysis and machine learning. He got his Ph.D. in electrical and computer engineering from the University of Iowa, USA. He has more than 70 patents and published more than 140 papers on leading IEEE Journals, conferences and workshops. He is a senior member of the IEEE. He is also an adjunct professor at School of Microelectronics, Fudan University, China.

DFX: Exploring the Design Space for Quality

Kaushik Narayanun
VP of Hardware Engineering, NVIDIA Corp.
2788 San Tomas Expy, Santa Clara, CA 95051, USA
knarayanun@nvidia.com

Keywords

Design quality, DFT, DFD, high-data volume, high-speed interfaces, programmable architectures, portability of test, AI, ML, RMA, EDA

Abstract

The ever-increasing demands of high-performance visual and accelerated computing has resulted in GPUs becoming some of the most complex ASICs being built today. The last few years have also seen an explosion in demand for unique silicon designs serving varied markets such as gaming, HPC, healthcare, smart cities, robotics and automotive. Process scaling is an important factor of delivering such continuous performance gains over the decades. Some of these designs push the limits of current chip manufacturing technology, growing to 80B transistors and beyond. Furthermore, these new designs are implemented with innovative new methods in physical design and are accelerated to reach the market at a staggering pace. Delivering outgoing quality in such an expeditious development environment presents unique test challenges related to test time, cost, power, advanced defectivity and diagnosability to list a few.

Current industry practice for structural testing of SOCs requires expensive automatic test equipment (ATE). As chip sizes grow exponentially, demand for memory per IO increases to meet low DPPM (defective-parts-per-million) requirements. In most cases, neither additional IOs are available for test, nor the speed of test paths through the IOs improve. With 2.5D and 3D chips becoming more mainstream, IOs available for test have further decreased. Scan compression schemes come at the cost of poor diagnosability or vendor specific design customizations thereby increasing the chip costs. This causes the test cost to increase drastically.

As chip volumes grows and 3D integration gets more adoption, it is critical to catch defects as early in the test flow as possible given the cost of waste that otherwise would accumulate. This makes ATE to system-level test (SLT) correlation one of the key factors, which has been historically a major challenge for test. In the small fraction of cases where it is possible, the cost of running structural tests on SLT environment is very expensive, which makes this process impractical. Universal EDA solutions for bridging the ATE-SLT correlation gap do not exist or are not practical.

As a new application to DFX design space, automotive and high-performance computing (HPC) markets require periodic in-field testing for safety and reliability. Structural tests provide extremely high coverage compared to functional patterns and most suited to satisfy the requirements of these segments. Existing schemes for in-field structural testing are limited due to long run times and/or low coverage. Additionally, the application of the structural patterns needs to be extremely secure to protect the confidential assets, which makes it even harder.

With the advent of sub-5nm transistor technologies, standard solutions that only look at tests' pass/fail results are not sufficient to catch marginal defects that are hard to detect. In the era of machine learning, we need look beyond test results and what human experts can see, into anomalous chip parameters in a sea of complex data.

DFX design space is expanding and rapidly changing, and it is not yet fully serviced by existing EDA solutions. Hence, current requirements push us to innovate across the spectrum of DFX architectures. This talk will provide a summary of various ideas used at NVIDIA to tackle these challenges: Using functional high-speed interfaces for test-data transfer, portable test solutions, programmable test architectures, improved design-for-debug features, emulation for DFX, and using machine learning to solve DFX problems.

Biography

Kaushik Narayanun is currently Vice President of Hardware and Head of DFX Engineering at NVIDIA. He received his B.E. degree in Electronics and Communication Engineering from University of Madras, M.S in Computer Engineering from University of California and an MBA from INSEAD. His career has been spent architecting and productizing DFX solutions for industry's leading SOCs and GPUs. His research focus is on reframing manfacturing test needs for unique markets such as automotive or AI datacenters, development of system level architectures that deliver improved outgoing quality and the use of machine intelligence to solve costly implementation workflows.

SiGe BICMOS Technology with Advanced Integration Solutions for mm-Wave and THz Applications

Dr.-Ing- Mehmet Kaynak
Leigniz Institute for High Performance Microelectronics (IHP)
Germany

Keywords

SiGe BiCMOS Technologies, Heterogeneous and 3D Integration, Wireless and broadband communication systems,

RFIC/MMIC, RF-MEMS

Abstract

In last decade, SiGe BiCMOS technologies open a new cost-efficient market first at mm-wave frequencies, then at sub-THz and THz range. Starting with the commercial use of automotive radars at 77 GHz, and the demand for 120/140 GHz radars, the market now has a strong interest on low-cost silicon based technologies for such high frequencies. The driving force behind the BiCMOS is not anymore only radar applications but also 5G/6G communication systems, operating at 60, 140 or even 240/300 GHz. Furthermore, the strong need on the driving circuitries of photonics components has created a mass market of high speed communication circuitries.

The latest developments on SiGe HBTs with fmax of beyond 700 GHz boosts the research and development effort on circuit and system area to take share from the new markets. In parallel to the developments on SiGe HBT performance, "More-than-Moore" path, which covers all the additional functionalities to the standard CMOS process (i.e. MEMS devices, microfluidics, photonics, etc…), allows to realize multi-functional circuits and systems.

Heterogeneous integration as a key and enabler technology for the multi-functional systems has also great importance these days. Hetero integration of different technologies such as high scaled CMOS, SiGe BiCMOS or even III-V ones has become real. Such integration of multi-chip technologies provides the highest performance with the most efficient size and power consumption; thus paves the way for next generation smart systems integration.

In this talk, approaches used for the integration of different More-than-Moore modules into a BiCMOS process will be presented. New hetero-integration technologies and corresponding challenges will be discussed

Biography

Dr. –Ing Mehmet Kaynak received his B.S degree from Electronics and Communication Engineering Department of Istanbul Technical University (ITU) in 2004, took the M.S degree from Microelectronic program of Sabanci University, Istanbul, Turkey in 2006 and received the PhD degree from Technical University of Berlin, Berlin Germany in 2014. He joined the technology group of IHP Microelectronics, Frankfurt (Oder), Germany in 2008. From 2008 to 2015, he has led the MEMS development at IHP. From 2015 to 2020, he had served as the the department head of Technology at IHP. Since 2020, he is continuing as a scientist at IHP Microelectronics.