1.1.1 Keynote: System Design Challenges for Next Generation Wireless and Embedded Systems
David Fuller - National Instruments, US

Application demands in our embedded world are growing dramatically. Consumer expectations and the industry's forward-looking technology roadmaps paint a picture of a connected world full of intelligent devices once thought to have fixed functionalities.

Researchers exploring next generation wireless systems, Internet of Things (IoT), and even machine-machined (M2M) communications face many challenges in making this vision a reality. Where once a single, isolated design flow addressed the discrete application, heterogeneous multi-processing architectures must be considered and embraced along with the connections to other devices and systems, and real-world sensor data.

As the systems grow in complexity, new design approaches must also be developed and employed to expedite the research, design, and development cycle.

David Fuller will outline challenges system designers face in developing cyber-physical systems and explore a graphical system design approach that includes hardware abstraction and comprehends a heterogeneous multiprocessor environment while embracing different models of computation. Through this new approach, system designers can shorten design cycles and the time to prototype ultimately accelerating deployment.

1.1.2 Keynote: The Growing Importance of Microelectronics from a Foundry Perspective
Gerd Teepe - Global Foundries, DE

Microelectronics is the dominant industrial technology of today. Its rate of innovation, spelled out by Moore's Law, is exceptional by any commercial metric, especially, as it has been on this trajectory for almost 40 years. It is not surprising, that other industrial sectors are taking advantage of the innovation engine of the semiconductor for its own product innovation: Cars are safer and more economic, medical diagnostics are performing to a significantly higher level, and energy efficiency from the generation to the consumer is a lot more efficient. "The Internet" has become the basis for our communication, organization and planning in our economies with significant impact to our society. However, the Semiconductor industry is under a powerful transformation marked by the following trends:

- Design Complexity is facing new challenges, as technological complexity is transferred to the design space at an accelerated pace
- The SOC is dominating the design space
- Intelligent Things are emerging with unprecedented cognitive and motion capabilities
- The supply chain transformation is in full motion, with the foundry model at the forefront

2.2 Panel: Emerging vs. Established Technologies: A Two Sphinxes’ Riddle at the Crossroads?
Organizer: Marco Casale-Rossi, Synopsys, Inc., US
Chair: Giovanni De Michiel, EPFL, CH

Panelists: Rob Atkin, Antun Domic, Manfred Horstmann, Robert Hum, Philippe Magarshack

Crossroads have always been challenging: they require a decision; in Egyptian and Greek mythology they were often guarded by two sphinxes trying to cheat the traveler with their riddles. The two sphinxes, the knight and the knave, the lady and the tiger, are just few instances of difficult puzzles that have kept logicians and mathematicians busy for the last 5,000 years. Today, you are walking down Moore's Law road when you come to a crossroads: one road brings you into the land of emerging technologies: 14, 10 and 7 nanometer, FDSOI, FinFET, 3D-IC, ... beyond and below; the other road holds you into the land of established technologies: 28, 40, 65, and 90 nanometers, possibly even above, A&M/S, MEMS,... Choosing the right road is critical to lead your project and your company to success, but making the right decision is increasingly difficult, as it encompasses complex technical and economic considerations. However, unlike the mythical traveler, you won't run into the sphinxes but, rather, into some of our industry best experts; unlike the sphinxes, they will strive to provide you with honest advice about the "road conditions", and you are allowed to ask multiple questions to them to figure out which road is the best for you.

2.3 Making Automotive Systems Safer and More Energy Efficient
Chair: Bart Vermeulen, NXP Semiconductors, NL; Martin Lukasiewycz, TUM CREATE, SG

Emulation-Based Robustness Assessment for Automotive Smart-Power ICs
Manuel Harnant, Thomas Nimaijer, Jérôme Kirschner, Christoph Grimm, Georg Peiz

In this paper we present a concept for assessing the robustness of automotive smart power ICs through lab measurements with respect to application variance and parameter spread. Classical compliance to the product specification, where only minimum and maximum values are defined, is not enough to assess device robustness since complex transients of application components cannot be defined within single specification parameters. That is why application fitness becomes a necessary task to reduce device failures, which may occur in the application. One solution would be to enhance traditional lab verification methods with a concept that considers application and parameter spread. This innovative concept is demonstrated on an electronic throttle control application. It has been emulated in real-time, including power amplification and application-relevant parameters. Monte Carlo experiments were carried out within the application space to evaluate the influence of parameter spread on selected system characteristics. Finally, an appropriate metric was used to quantify the robustness of the micro-electronic device within its application.

Keywords: Automotive Power Micro-electronics, Electronic Throttle Control, Post-Silicon Verification, Application Fitness, Worst-Case Distance

Startup Error Detection and Containment to Improve the Robustness of Hybrid FlexRay Networks
The development and research on in-vehicle networks (IVNs) is driven by two main requirements: bandwidth and robustness. In this paper, we address the robustness requirement. We focus on FlexRay IVNs that are used for safety-critical applications. We analyze and discuss faults that may affect the startup and operation of a FlexRay network. These failures may not only occur during the startup phase of the vehicle, but they may also happen due to a bus problem that requires the bus to be reinitialized during normal operation. Here, any startup failure leads to a critical situation like a brake system failure. The fault scenarios we discuss in this paper are the resetting leading coldstart node (RLCN), the deaf coldstart node (DCN), and the babbling idiot (BI). These faults are described in literature, but neither precise behavior of all involved nodes, nor a clear solution is provided to contain their impact. The idea of a bus guardian (BG) is given in a draft specification of the FlexRay consortium, but no details are given. In this paper, we extend on these ideas by investigating and implementing a detailed (BG) concept, based on our own analysis. We subsequently evaluate the successful containment of the three fault types in simulation. We also quantify the chip area cost of our solution.

Emerging hierarchical design methodologies based on the use of Pareto-optimal fronts (PoFs) are promising candidates to reduce the bottleneck in the design of analog circuits. However, they are still challenging to efficiently evaluate performance bound of high-precision analog circuits with multiple parameter variations at nano-scale. In this paper, we present a method to analyze the transient error resiliency of a digital circuit. This method is based on an analytical model. It models a transient error as a random function and finds the vulnerable number of bits for each node. We perform a case study on a circuit implementation of a well-known adaptive filter algorithm. The results from the analytical and simulation models show that the analytical model is accurate enough to estimate the effects of transient errors on the performance of a digital circuit. Our analytical method also reduces the analysis time significantly in a design phase.

This paper describes an electromigration-aware and IR-Drop avoidance routing approach considering multiport multiterminal (MP/MT) signal nets of analog integrated circuits (IC). The design phase is therefore essential to evaluate, and where necessary improve, the resilience of a circuit to all possible transient errors. In this paper, we present a method to analyze the transient error resiliency of a digital circuit. This method is based on an analytical model. It models a transient error as a random function and finds the vulnerable number of bits for each node. We perform a case study on a circuit implementation of a well-known adaptive filter algorithm. The results from the analytical and simulation models show that the analytical model is accurate enough to estimate the effects of transient errors on the performance of a digital circuit. Our analytical method also reduces the analysis time significantly in a design phase.

2.4 Modern Challenges in Analog and Mixed-Signal Design

Chair: Georges Gielen, KU Leuven, BE; Günther Dündar, Bogazici University, TR

(18) Electromigration-Aware and IR-Drop Avoidance Routing in Analog Multiport Terminal Structures
Ricardo Martins, Nuno Lourenço, António Canelas, Nuno Horta
This paper describes an electromigration-aware and IR-Drop avoidance routing approach considering multiport multiterminal (MP/MT) signal nets of analog integrated circuits (IC). The effects of current densities and temperature on the interconnects may cause the malfunction/failure of a circuit due to IR-Drop or electromigration (EM). These become increasingly more relevant with the ongoing reduction of circuit sizes caused by the evolution of the nanoscale integration processes. Therefore, EM and IR-Drop effects must be taken into account in the design of both power networks and signal wires of analog and mixed-signal ICs, to make their impact on the circuits’ reliability negligible. In previous EM and IR-Drop-aware analog IC routing approaches, "dot-models" are assumed for the terminals, i.e., each terminal has only one port that need to be routed. However, in practice, analog standard cells usually contain multiple electrically-equivalent locations, often distributed over different fabrications layers, where legal connections can be made, i.e., MP terminals, which need to be properly explored. The design flow is detailed, and the applicability of the approach is demonstrated with experimental results, and also, by generating the routing of an analog circuit structure for the UMC 130nm design process.

(19) Zonotope-based Nonlinear Model Order Reduction for Fast Performance Bound Analysis of Analog Circuits with Multiple-interval-valued Parameter Variations
Yang Song, Sai Manoj P.D., Hao Yu
It is challenging to efficiently evaluate performance bound of high-precision analog circuits with multiple parameter variations at nano-scale. In this paper, a nonlinear model order reduction is proposed to deploy zonotope-based model for multiple-interval-valued parameter variations. As such, one can have a zonotope-based reachability analysis to generate a set of trajectories with performance bound defined. By further constructing local parameterized subspaces to approximate a number of zonotopes along the set of trajectories, one can perform nonlinear model order reduction to generate the performance bound under parameter variations. As shown by numerical experiments, the zonotope-based nonlinear macromodeling by order of 19 achieves up to 50X speedup when compared to Monte Carlo simulations of the original model, and up to 50% smaller error when compared to previous parameterized nonlinear macromodeling under the same order.

(20) Implementation Issues in the Hierarchical Composition of Performance Models of Analog Circuits
M. Velasco-Jiménez, R. Castro-López, E. Roa, F.V. Fernández
Emerging hierarchical design methodologies based on the use of Pareto-optimal fronts (PoFs) are promising candidates to reduce the bottleneck in the design of analog circuits. However, little work has been reported about how to transmit the information provided by the PoF to low level hierarchical blocks through the hierarchy to compose the performance models of higher-level blocks. This composition actually poses several problems such as the dependence of the PoF performances on the surrounding circuitry and the complexity of dealing with...
2.5 Low-Power and Efficient Architectures

Chairs: Cristina Sikano, Politecnico di Milano, IT; Todd Austin, University of Michigan, US

[27] Energy-Efficient Scheduling for Memory-Intensive GPGPU Workloads
Seokwo Sang, Minseok Lee, John Kim, Woong Seo, Yeongon Cho, Sookyung Ryu
High performance for a GPGPU workload is obtained by maximizing parallelism and fully utilizing the available resources. However, this is not necessarily energy efficient, especially for memory-intensive GPGPU workloads. In this work, we propose Throttle CTA (cooperative-thread array) Scheduling (TCS) where we leverage two type of throttling - throttling the number of active cores and throttling of warp execution in the cores - to improve energy-efficiency for memory-intensive GPGPU workloads. The algorithm requires the global CTA or thread block scheduler to reduce the number of cores with assigned thread blocks while leveraging the local warp scheduler to throttle memory requests for some of the cores to further reduce power consumption. The proposed TCS scheduling does not require off-line analysis but can be done dynamically during execution. Instead of relying on conventional metrics such as miss-per-kilo-instruction (MPKI), we leverage the memory access latency metric to determine the memory intensity of the workloads. Our evaluations show that TCS reduces energy by up to 48% (38% on average) across different memory-intensive workload while having very little impact on performance for compute-intensive workloads.

[28] eCVM: Energy-Efficient Distributed Scratchpad Video Memory Architecture for the Next-Generation High Efficiency Video Coding
Felipe Sampaio, Muhammad Shafique, Bruno Zatt, Siegko Bampi, Jörg Henkel
An energy-efficient distributed Scratchpad Video Memory Architecture (dSVM) for the next-generation parallel High Efficiency Video Coding is presented. Our dSVM combines private and overlapping (shared) Scratchpad Memories (SPMs) to support data reuse within and across different cores concurrently executing multiple parallel HEVC threads. We developed a statistical method to size and design the organization of the SPMs along with a supporting memory reading policy for energy efficiency. The key is to leverage the HEVC and video content knowledge. Furthermore, we integrate an adaptive power management policy for SPMs to manage the power states of different memory parts at run time depending upon the varying video content properties. Our experimental results illustrate that our dSVM architecture reduces the overall memory energy consumption by up to 51%-61% compared to parallelized state-of-the-art solutions [1, 11]. The dSVM external memory energy savings increase with an increasing number of parallel HEVC threads and size of search window. Moreover, our SPM power management reacts to the current video properties and achieves up to 54% on-chip leakage energy savings.

**Keywords** - Video Memory, Scratchpad Memory, HEVC, Application-Specific Optimizations, Energy Efficiency, Adaptivity.

---

Feasibility Exploration of NVM Based I-Cache through MDH Enhancements
Manu Komalan, José Ignacio Gómez Pérez, Christian Tenllado, Praveen Raghavan, Matthias Hartmann, Francky Cathoor

SRAM-based memory systems are plagued by a number of problems like sub-threshold leakage and susceptibility to read/write failure with dynamic voltage scaling schemes or low supply voltage. Non-Volatile Memory (NVM) technologies are being explored extensively nowadays to replace the conventional SRAM memories even for level 1 (L1) caches. These NVMs like Spin Torque Transfer RAM (STT-MRAM), Resistive RAM (ReRAM) and Phase Change RAM (PRAM) are less hindered by leakage problems with technology scaling and consume lesser area. However, simple replacement of SRAM by NVMs is not a viable option due to their write related issues. The main focus of this paper is the exploration of write delay and write energy issues in a NVM based L1 instruction cache (I-cache) for an ARM like single core system. We propose a NVM I-cache and extend its MSHR (Miss Status Handling Register) functionality to address the NVMs write related issues. According to our simulations, appropriate tuning of selectivity architecture parameters can reduce the performance penalty introduced by the NVM (-45%) to extremely tolerable levels (-1%) and show energy gains up to 35%. Furthermore, on configuring our modified NVM based system to occupy area comparable to the original SRAM-based configuration, it outperforms the SRAM baseline and leads to even more energy savings.

---

EVX: Vector Execution on Low Power EDGE Cores
Milovan Dutc, Oscar Palomar, Aaron Smith, Osman Unsal, Adrian Cristal, Mateo Valero, Doug Burger

In this paper, we present a vector execution model that provides the advantages of vector processors on low power, general purpose cores, with limited additional hardware. While accelerating data-level parallel (DLP) workloads, the vector model increases the efficiency and hardware resources utilization. We use a modest dual issue core based on an Explicit Data Graph Execution (EDGE) architecture to implement our approach, called EVX. Unlike most DLP accelerators which utilize additional hardware and increase the complexity of lowpower processors, EVX leverages the available resources of EDGE cores, and with minimal costs allows for specialization of the resources. EVX adds a control logic that increases the core area by 2.1%. We show that EVX yields an average speedup of 3x compared to a scalar baseline and outperforms multimedia SIMD extensions.

---

Program Affinity Performance Models for Performance and Utilization
Ryan W. Moore and Bruce R. Childers

Multithreaded applications have a wide variety of behavior, causing complex interactions with today's chip multiprocessor machines. Application threads may have large private working sets, and may compete for cache space and memory bandwidth. These threads benefit from large private caches. Other threads may share data or communicate, and thus, execute more quickly if using shared caches. Many applications fall somewhere in between, requiring careful thread-to-core assignments to maximize performance. Yet because of the large number of thread-to-core assignments on today's chip multiprocessors, it is time and energy prohibitive to exhaustively try and determine the best assignment. In this paper, we present and demonstrate application performance models that predict application performance given a proposed thread-to-core assignment. We show how these models can be quickly built and used to select thread-to-core assignments for multiple programs and to improve system utilization.

---

Advanced SIMD: Extending the Reach of Contemporary SIMD Architectures
Matthias Boettcher, Bashir M. Al-Hashimi, Mbou Eyole, Giacomo Gabriei, Alastair Reid

SIMD extensions have gained widespread acceptance in modern microprocessors as a way to exploit data-level parallelism in general-purpose cores. Popular SIMD architectures (e.g. Intel SSE/AVX) have evolved by adding support for wider registers and data paths, and advanced features like indexed memory accesses, per-lane predication and inter-lane instructions, at the cost of additional silicon area and design complexity. This paper evaluates the performance impact of such advanced features on a set of workloads considered hard to vectorize for traditional SIMD architectures. Their sensitivity to the most relevant design parameters (e.g. register/datapath width and L1 data cache configuration) is quantified and discussed. We developed an ARMv7 NEON based ISA extension (ARGON), augmented a cycle accurate simulation framework for it, and derived a set of benchmarks from the Berkeley dwarfs. Our analyses demonstrate how ARGON can, depending on the structure of an algorithm, achieve speedups of 1.5x to 16x.

---

A Tightly-coupled Hardware Controller to Improve Scalability and Programmability of Shared-Memory Heterogeneous Clusters
Paolo Burgio, Robin Danilo, Andrea Marongiu, Philippe Cousy, Luca Berti

Modern designs for embedded many-core systems increasingly include application specific units to accelerate key computational kernels with orders-of-magnitude higher execution speed and energy efficiency compared to software counterparts. A promising architectural template is based on heterogeneous clusters, where simple RISC cores and specialized HW units (HWPU) communicate in a tightly-coupled manner via L1 shared memory. Efficiently integrating processors and a high number of HW Processing Units (HWPU) in such a system poses two main challenges, namely, architectural scalability and programmability. In this paper we describe an optimized Data Pump (DP) which connects several accelerators to a restricted set of communication ports, and acts as a virtualization layer for programming, exposing FIFO queues to offload "HW tasks" to them through a set of lightweight APIs. In this work, we aim at optimizing both these mechanisms, for respectively reducing modules area and making programming sequence easier and lighter.

---

2.6 Real-Time Memory Hierarchies

Chairs: Benny Aksesson, CTU Prague, CZ; Giuseppe Lipari, ENS - Cachan, FR

---

On the Correctness, Optimality and Precision of Static Probabilistic Timing Analysis
Sebastian Altmeier, Robert T. Davis

In this paper, we investigate Static Probabilistic Timing Analysis (SPTA) for single processor systems that use a cache with an evict-on-miss random replacement policy. We show that previously published formulae for the probability of a cache hit can produce results that are optimistic and unsound when used to compute probabilistic Worst-Case Execution Time (pWCET) distributions. We investigate the correctness, optimality, and precision of different approaches to SPTA. We prove that one of the previously published formulae for the probability of a cache hit is optimal with respect to the limited information that it uses. We improve upon this formulation by using extra information about cache contention. To investigate the precision of our new approach, we use SPTA, we introduce a simple hardware method that computes a precise pWCET distribution, albeit at the cost of exponential complexity. Further, we integrate this precise approach, applied to small numbers of frequently accessed memory blocks, with imprecise analysis of other memory blocks, to form a combined approach that improves precision, without significantly increasing its complexity. The performance of the various approaches is compared on benchmark programs.

---

WGET-Centric Dynamic Instruction Cache Locking
Huping Ding, Yun Liang, Tulika Mitra

Cache locking is an effective technique to improve timing predictability in real-time systems. In static cache locking, the locked memory blocks remain unchanged throughout the program execution. Thus static locking may not be effective for large programs where multiple memory blocks are competing for few cache lines available for locking. In comparison, dynamic cache locking overcomes cache space limitation through time-multiplexing of locked memory blocks. Prior dynamic locking technique partitions the program into regions and takes independent locking decisions for each region. We propose a flexible loop-based dynamic cache locking approach. We not only select the memory blocks to be locked but also the locking points (e.g., loop level). We judiciously allow memory blocks from the same loop to be locked at different program points for WGET improvement. We design a constraint-based approach that
Minimizing Stack Memory for Hard Real-time Applications on Multicore Platforms
Chuansheng Dong, Halbo Zeng
Multicore platforms are increasingly used in realtime embedded applications. In the development of such applications, an efficient use of RAM memory is as important as the effective scheduling of the application tasks. Preemption Threshold Scheduling is a well-known technique for controlling the degree of preemption, possibly improving system schedulability, and allowing savings in stack space. In this paper, we target at the optimal mapping of tasks to cores and the assignment of the scheduling parameters for systems scheduled with preemption thresholds. We formulate the optimization problems using Mixed Integer Linear Programming framework, and propose an efficient heuristic as an alternative. We demonstrate the efficiency and quality of both approaches with extensive experiments using random systems as well as two industrial case studies.

Time-predictable Execution of Multithreaded Applications on Multicore Systems
Ahmed Alhammad, Rodolfo Pellizzoni
In multicore systems, contention for access to main memory between application threads complicates timing analysis and may lead to pessimistic bounds on execution time. This is particularly problematic for real-time applications, which require provable bounds on worst-case performance. In this work, we employ a predictable execution model to schedule memory accesses performed by application threads without relying on unpredictable hardware arbiters. In addition, we statically schedule application's threads with the objective to minimize the application's makespan. Our experimental evaluation with NAS Parallel Benchmarks on 4-core system indicates that the proposed execution scheme yields an aggregated improvement of 21% over contention execution in which application's threads uncontrollably access main memory.

2.7 Yield and Reliability for Robust Systems
Chairs: Joan Figueras, UPC, ES; Jose Pineda de Gyerex, NXP, NL

Approximating the Age of RF/Analog Circuits through Re-characterization and Statistical Estimation
Doochang Chang, Sule Ozay, Ozgur Sinanoglu, Ramesh Karri

Comprehensive Analysis of Alpha and Neutron Particle-induced Soft Errors in an Embedded Processor at Nanoscales
Mohja Ebrahimi, Adrian Evans, Mehrb B. Tahoori, Razi Seyyedi, Enrico Costenaro, Dan Alexandrescu
Radiation-induced soft errors have become a key challenge in advanced commercial electronic components and systems. We present results of Soft Error Rate (SER) analysis of an embedded processor. Our SER analysis platform accurately models all generation, propagation and masking effects starting from a technology response model derived using TCAD simulations of the device level all the way to application masking. The platform employs a combination of device level, analytical error propagation at logic level and fault emulation at the architecture/application level to provide the detailed contribution of each component (flip-flops, combinational gates, and SRAMs) to the overall SER. At each stage in the modeling hierarchy, an appropriate level of abstraction is used to propagate the effect of errors to the next higher level. Unlike previous studies which are based on very simple test chips, analyzing the entire processor gives more insight into the contributions of different components to the overall SER. The results of this analysis can assist circuit designers to adopt effective hardening techniques to reduce the overall SER while meeting required power and performance constraints.

Bias Temperature Instability Analysis of FinFET Based SRAM Cells
Seyab Khan, Innocent Agbo, Said Hamdiou, Hal Kukner, Ben Kaszer, Praveen Raghavan, Franky Catthoor
Bias Temperature Instability (BTI) is posing a major reliability challenge for today's and future semiconductor devices as it degrades their performance. This paper provides a comprehensive BTI impact analysis, in terms of time-dependent degradation, of FinFET based SRAM cell. The evaluation metrics are read Static Noise Margin (SNM), hold SNM and Write Trip Point (WTP); while the aspects investigated include BTI impact dependence on the supply voltage, cell strength, and design styles (6 versus 8 Transistors cell). A comparison between FinFET and planar CMOS based SRAM cells degradation is also covered. The simulation performed on FinFET based cells for 108 seconds of operation under nominal Vdd show that Read SNM degradation is 16.72%, which is 1.17x faster than hold SNM and 1.21x faster than hold SNM, while WTP improves by 6.82%. In addition, a supply voltage increment of 25% reduces the Read SNM degradation by 40%, while strengthening the cell pull-down transistors by 1.5x reduces the degradation by only 22%. Moreover, the results reveal that 8T cell degrades 1.31X faster than 6T cell, and that FinFET cells are more vulnerable (~2x) to BTI degradation than planar CMOS cells.

SSFB: A Highly-Efficient and Scalable Simulation Reduction Technique for SRAM Yield Analysis
Manish Rana, Ramon Canal
Estimating extremely low SRAM failure probabilities by conventional Monte Carlo (MC) approach requires hundreds-of-thousands simulations making it an impractical approach. To alleviate this problem, failure-probability estimation methods with a smaller number of simulations have recently been proposed, most notably variants of consecutive mean-shift based Importance Sampling (IS). In this method, a large amount of time is spent simulating data points that will eventually be discarded in favor of other data points with minimum norm. This can potentially increase the simulation time by orders of magnitude. To solve this very important limitation, in this paper, we introduce SSFB: a novel SRAM failure-probability estimation method that has much better cognition of the data points compared to conventional approaches. The proposed method starts with radial simulation of a single point and reduces discarded simulations by: a) random sampling only when it reaches a failure boundary and after that continues again with radial simulation of a chosen point, and b) random sampling is performed only within a specific failure-range which decreases in each iteration. The proposed method is also scalable to higher dimensions (more input variables) as sampling is done on the surface of the hyper-sphere, rather than within the hypersphere as other techniques do. Our results show that using our method we can achieve an overall 40x reduction in simulations compared to consecutive mean-shift IS methods while remaining within the 0.01-Sigma accuracy.

INFORMER: An Integrated Framework for Early-Stage Memory Robustness Analysis
Shikinath Ganapathy, Ramon Canal, Dan Alexandrescu, Enrico Costenaro, Antonio Gonzalez, Antonio Rubio
With the growing importance of parametric (process and environmental) variations in advanced technologies, it has become a serious challenge to design reliable, fast and low-power embedded memories. Adopting a variation-aware design paradigm requires a holistic perspective of memory-wide metrics such as yield, power and performance. However, accurate estimation of such metrics is largely dependent on circuit implementation styles, technology parameters and architecture-level specific variants. In this paper, we propose a fully automated tool - INFORMER - that helps high-level designers estimate memory reliability metrics rapidly and accurately. The tool relies on accurate circuit-level simulations of failure mechanisms such as soft-errors and parametric failures. The statistics obtained can then help couples low-level metrics with higher-level design choices. A new technique for rapid estimation of low-probability failure events is also proposed. We present three use-cases of our prototype tool to demonstrate its diverse capabilities in autonomously guiding large SRAM based robust memory designs.

Wear-out Analysis of Error Correction Techniques in Phase-change Memory
Caiho Heftman, Luis Ramos, Rodolfo Azevedo, Guido Araujo
Phase-Change Memory (PCM) is new memory technology and a possible replacement for DRAM, whose scaling limitations require new lithography technologies. Despite being promising, PCM has limited endurance (its cells withstand roughly 108 bit-flips before failing), which prompted the adoption of Error Correction Techniques (ECTs). However, previous lifetime analyses of ECTs did not consider the difference between the bit-flip frequencies of data and code bits, which may lead to inaccurate wear-out analyses for the ECTs. In this work, we improve the wear-out analysis of PCM by modeling and analyzing the bit-flip probabilities of five ECTs. Our models also enable an accurate estimation of energy consumption and analysis of the endurance-energy trade-off for each ECT.

Approximating the Age of RF/Analog Circuits through Re-characterization and Statistical Estimation
Doochang Chang, Sule Ozay, Ozgur Sinanoglu, Ramesh Karri
Counterfeit ICs have become an issue for semiconductor manufacturers due to impacts on their reputation and lost revenue. Counterfeit ICs are either products that are intentionally mislabeled or legitimate products that are extracted from electronic waste. The former is easier to detect whereas the latter is harder since they are identical to new devices but display degraded performance due to environmental and use stress conditions. Detecting counterfeit ICs that are extracted from electronic waste requires an approach that can approximate the age of manufactured devices based on their parameters. In this paper, we present a methodology that uses information on both fresh and aged ICs and tries to distinguish between the fresh and aged population based on an estimate of the age. Since analog devices age mainly due to their bias stress, input signals play less of a role. Hence, it is possible to use simulation models to approximate the aging process, which would give us access to a large population of aged devices. Using this information, we can construct a statistical model that approximates the age of a given circuit. We use a Low noise amplifier (LNA) and an NMOS LC oscillator to demonstrate that individual aged devices can be accurately classified using the proposed method.

2.8 Hot Topic: Technology Transfer towards Horizon 2020
Organizer: Rainer Leupers, RWTH Aachen, Germany
Chair: Norbert Wehn, TU Kaiserslautern, Germany

European research projects produce many excellent results, and the quality of research papers at DATE and other major European conferences is often outstanding. But how many academic research results in computing technologies and EDA actually make it into industrial practice? In the context of the transition into the Horizon 2020 framework program, the European research community is currently investigating novel ways of stimulating additional academia-industry technology transfer. This special session contributes by discussing concrete transfer experiences and new concepts. Furthermore it will exemplify several success stories from both academic and industrial perspectives.

3.2 Panel: The World is Going... Analog & Mixed-Signal! What about EDA?
Organizer: Marco Casale-Rossi, Synopsys, Inc., US.
Chair: Pietro Palella, STMicroelectronics, Italy

Contrary to a common belief, the world is not going digital! Analog and mixed-signal electronics is more and more important and yet pervasive. This is due both to the increasing systems integration, by nature leading to heterogeneity, and to the complex, digital computing functions being complemented by scores of on-chip analog functions, interacting with people, environment, and other systems. Specialty silicon foundries are now stable members of top ten revenue rankings. This technology trend demands for more design automation in both implementation and verification domains. Lossless interfaces between digital and analog design environments, multi-technology support, mixed-signal simulation engines - but also debugging aids - are no longer a nice to have. According to IBS [1], the cost of implementing and verifying the mixed-signal functions is generally over 50% of the design costs even though the mixed-signal transistors can be as low as 3% of the total. What are the critical requirements, moving forward, and what is EDA industry doing to serve the needs of this increasingly important semiconductor industry segment?

3.3 Secure Hardware Primitives and Implementations
Chair: Paolo Maselli, TUM, FR, Patrick Schaumont, Virginia Tech, US

Lightweight Code-based Cryptography: QC-MDPC McEliece Encryption on Reconfigurable Devices
Ingo von Maauch, Tim Guneysu

With the break of RSA and ECC cryptosystems in an era of quantum computing, asymmetric code-based cryptography is an established alternative that can be a potential replacement. A major drawback are large keys in the range between 50 kByte to several MBbyte that prevented real-world applications of code-based cryptosystems so far. A recent proposal by Misoczki et al. showed that quasi-cyclic moderate density parity-check (QC-MDPC) codes can be used in McEliece encryption - reducing the public key to just 0.6 kByte to achieve a 80-bit security level. Despite of reasonably small key sizes that could also enable small designs, previous work only report high-performance implementations with high resource consumptions of more than 13,000 slices on a large Xilinx Virtex-6 FPGA for a combined en-/decryption unit. In this work we focus on lightweight implementations of code-based cryptography and demonstrate that McEliece encryption using QC-MDPC codes can be implemented with a significantly smaller resource footprint - still achieving reasonable performance sufficient for many applications, e.g., challenge-response protocols or hybrid firmware encryption. More precisely, our design requires just 68 slices for the encryption and around 150 slices for the decryption unit and is able to en-/decrypt an input block in 2.2 ms and 13.4 ms, respectively.

On the Assumption of Mutual Independence of Jitter Realizations in P-Trng Stochastic Models
Patrick Haddad, Yannick Teglia, Florent Bernard, Viktor Fischer

Security in true random number generation in cryptography is based on entropy per bit at the generator output. The entropy is evaluated using stochastic models. Several recent works propose stochastic models based on assumptions related to selected physical analog phenomena such as noise or jittery signal and on the knowledge of the principle of randomness extraction from the obtained analog signal. However, these assumptions simplify considerably the underlying analog processes, which include several noise sources. In this paper, we present a new comprehensive multilevel approach, which enables to build the stochastic model based on detailed analysis of noise sources starting at transistor level and on conversion of the noise to the clock jitter exploited at the generator level. Using this approach, we can estimate proportion of the jitter coming only from the thermal noise, which is included in the total clock jitter.

Clock-Modulation Based Watermark for Protection of Embedded Processors
Jedrzej Kufel, Peter Wilson, Stephen HW, Bashir M. Al-Hashimi, Paul N. Whatmough, James Myers

This paper presents a novel watermark generation technique for the protection of embedded processors. In previous work, a load circuit is used to generate detectable watermark patterns in the ASIC power supply. This approach leads to hardware area overheads. We propose removing the dedicated load circuit entirely, instead to compensate the reduced power consumption the watermark pattern is emulated by reusing existing clock gated sequential logic as a zero-overhead load circuit and modulating the clock gating enable signal with the watermark sequence. The proposed technique has been validated through experiments using two ASICs in 65nm CMOS, one with an ARM Cortex-M0 microcontroller and one with a Cortex-A5 microprocessor. Silicon measurement results verify the viability of the technique for embedded processors. Furthermore, the proposed clock modulation technique demonstrates a significant area reduction, without compromising the detection performance. In our experiments an area overhead reduction of 98% was achieved. Through reuse of existing logic and reduction of watermark hardware implementation costs, the proposed clock modulation technique offers an improved robustness against removal attacks.

Index Terms - Watermarking, Embedded Systems, CPA

3.4 Modeling and Optimization of Power Distribution Networks
Chair: Luca Daniel, MIT, US; Stefano Givet-Talocia, Politecnico di Torino, Italy

Sensitivity-based Weighting for Passivity Enforcement of Linear Macromodels in Power Integrity Applications
A. Liboli, S. Givet-Talocia, M. Bandini, A. China

The electrical performance of Power Distribution Networks (PDNs) is usually assessed by computing frequency responses through quasi-static or full-wave electromagnetic solvers. Such
Characterizing Power Delivery Systems with On/Off-Chip Voltage Regulators for Many-Core Processors
Xuan Wang, Jiang Xu, Zhe Wang, Kevin J. Chen, Xiaowen Wu, Zhehui Wang
Design of power delivery system has great influence on the power management in many-core processor systems. Moving voltage regulators from off-chip to on-chip gains more and more interest in the power delivery system design, because it is able to provide fast voltage scaling and multiple power domains. Previous works are proposed to implement power efficient on-chip regulators. It is also important to analyze the characteristics of the entire power delivery system to explore the tradeoff between the promising properties and costs of employing on-chip regulators. In this work, we develop an analytical model to evaluate important characteristics of the power delivery system, including on-chip/off-chip voltage regulators and the passive on-chip on-board parasitic. Compared with SPICE simulations, our model achieves a fast system-level evaluation with comparable accuracy. Based on the model, geometric programming is utilized to find the optimal power efficiency of different architectures of power delivery systems under constraints of output voltage stability and area. Experiments show that compared with the conventional architecture using off-chip regulators, the hybrid one using both on-chip and off-chip voltage regulators achieves 1.0% power efficiency improvement and 68% area reduction of voltage regulators on average. We conclude that the hybrid architecture has potential for high power efficiency and small area at heavy workload, but careful account for the overhead of on-chip regulators is needed.

Cost-Effective Decap Selection for Beyond Die Power Integrity
Yi-En Chen, Tu-Hsiung Tsai, Shi-Hao Chen, Hung-Ming Chen
In designing reliable power distribution networks (PDN) for power integrity (PI), it is essential to stabilize voltage supply to devices on chip. We usually employ decoupling capacitor (decap) to suppress the noise generated by the switching of devices. There have been numerous prior works on how to select/inset decaps in chip, package, or board to maintain PI, however optimal decap selection is usually not applicable due to design budget and manufacturability. Moreover, design cost is seldom touched or mentioned. In this research, we propose an efficient methodology “PDC-PSO” to automatically optimizing the selection of available decaps. This algorithm not only takes advantage of particle swarm optimization (PSO) to stochastically search the design space, but takes the most effective range of decaps into consideration to outperform the basic PSO. We apply this to three real package designs and the results show that, compared to the original decap selection by rules of thumb, our approach could shorten the design period and we have better combination of decaps at the same or lower cost. In addition, our methodology can also consider package-board co-design in optimizing different operation frequencies.

Efficient Analysis of Variability Impact on Interconnect Lines and Resistor Networks
Jorge Fernández Villena, L. Miguel Silveira
Continued technology scaling coupled with limited lithographic capabilities is a leading cause of increased design variability. In the nanometer regime lithography tools have failed to keep pace with Moore’s Law and printed feature sizes are a small fraction of the wavelength of light used in current processes. Such sub-wavelength printing makes features highly susceptible to perturbations in the lithographic process conditions which leads to printed designs exhibiting increased variability. Such variability directly affects design behavior and performance in multiple ways. One of the areas of concern is power grid (PG) design, where lithographic errors may locally modify the wire widths. These variations, that may affect any and all wires in the grid, have a critical impact on the power distribution across the chip, introducing considerable current fluctuations which are a potential cause for electromigration effects. To analyze and account for the impact of these errors requires a complete extraction of the PG, which generates a large resistive network, potentially with several million elements, whose simulation is computationally challenging. This paper proposes a fast and accurate variability analysis of very large resistor networks, such as PG extracted netlists, that allows estimating the effects of multiple parameter settings in reasonable time. The proposed model can be easily combined with Litho/CMP simulators in order to boost much needed design-aware lithography.

Implicit Index-aware Model Order Reduction for RLC/RC Networks
Nicodemus Banagaya, Giuseppe AI, Wil. H.A. Schilders, Caren Tischendorf
This paper introduces the implicit-IMOR method for differential algebraic equations. This method is a modification of the Index-aware model order reduction (IMOR) method proposed in our earlier papers which is the explicit-IMOR method. It also involves first splitting the differential-algebraic equations (DAEs) into differential and algebraic parts using a basis of projectors. In contrast with the explicit-IMOR method, the implicit-IMOR method leads to implicit and algebraic parts. We demonstrate the implicit-IMOR method using the RLC/RC networks, but it can also be applied to other problems which lead to differential-algebraic equations.

Implicit-Index-Aware Model Order Reduction for RLC/RC Networks
Nicodemus Banagaya, Giuseppe AI, Wil. H.A. Schilders, Caren Tischendorf
This paper introduces the implicit-IMOR method for differential algebraic equations. This method is a modification of the Index-aware model order reduction (IMOR) method proposed in our earlier papers which is the explicit-IMOR method. It also involves first splitting the differential-algebraic equations (DAEs) into differential and algebraic parts using a basis of projectors. In contrast with the explicit-IMOR method, the implicit-IMOR method leads to implicit and algebraic parts. We demonstrate the implicit-IMOR method using the RLC/RC networks, but it can also be applied to other problems which lead to differential-algebraic equations.
Existing memory subsystems and TDM NoCs for real-time systems are optimized independently in terms of cost and performance by configuring their arbiter according to the bandwidth and/or latency requirements of their clients. However, when they are used in conjunction, and run in different clock domains, i.e. they are decoupled, there exists no structured methodology to select the NoC interface width and operating frequency for minimizing area and/or power consumption. Moreover, the multiple arbitration points, one in the NoC and the other in the memory subsystem, introduce additional overhead in the worst-case guaranteed latency. These make it hard to design cost-efficient real-time systems. The three main contributions in this paper are: (1) We present a novel methodology to couple any existing TDM NoC with a real-time memory controller and compute the different NoC interface width and operating frequency combinations for minimal area and/or power consumption. (2) For two different TDM NoC types, one a packet-switched and the other circuit-switched, we show the trade-off between area and power consumption with the different NoC configurations, for different DRAM generations. (3) We compare the coupled and decoupled architectures with the two NoCs, in terms of guaranteed worst-case latency, area and power consumption by synthesizing the designs in 40 nm technology. Our experiments show that using a coupled architecture in a system consisting of 16 clients results in savings of over 44% in guaranteed latency, 18% and 17% in area, 19% and 11% in power consumption for a packet-switched and a circuit-switched TDM NoC, respectively, with different DRAM types.

**Cyber Physical Systems: Security and Co-design**

Javier Jalle, Leonidas Kosmidis, Jaume Abella, Eduardo Quitones, Francisco J. Cazorla

Probabilistic Timing Analysis (PTA) reduces the amount of information needed to provide tight WCET estimates in real-time systems with respect to classic timing analysis. PTA imposes new requirements on hardware design that have been shown implementable for single-core architectures. However, no support has been proposed for multicore so far. In this paper, we propose several probabilistically-analyzable bus designs for multicore processors ranging from 4 cores connected with a single bus, to 16 cores deploying a hierarchical bus design. We derive analytical models of the probabilistic timing behaviour for the different bus designs, show their suitability for PTA and evaluate their hardware cost. Our results show that the proposed bus designs (i) fulfill PTA requirements, (ii) allow deriving WCET estimates with the same cost and complexity as in single-core processors, and (iii) provide higher guaranteed performance than single-core processors, 3.4x and 6.6x on average for an 8-core and a 16-core setup respectively.

**Programmable Decoder and Shadow Threads: Tolerate Remote Code Injection Exploits with Diversified Redundancy**

Zyi Liu, Wedong Shi, Shouhui Xu, Zhiqiang Lin

We present a lightweight hardware framework for providing high assurance detection and prevention of code injection attacks using a lockstep diversified shadow execution. Recent studies show that hardware diversification can detect software attacks by checking the consistency of their behavior simultaneously. Unfortunately, the severe performance degradation and extra system costs caused by these methods are unacceptable in many applications. This paper presents a hardware-level, lockstep shadow thread framework to enrich the diversity of the software execution, with the facilitation from programmable hardware decoder and novel CPU support of tightly coupled shadow thread technique. Specifically, given a piece of (legacy) binary code, we first generate diversified binary versions using an offline binary rewriter and programmable hardware binary translator at runtime. Two diversified binary code images are launched as dual simultaneous threads in the hardware layer with one as the primary thread and the other one as shadow thread. Instructions from the shadow thread are not executed but just compared, and thus incur no OS side-effects. The extended CPU is able to decode instructions from both threads, and dispatch them to the next stage pipeline for a lockstep comparison. Any mismatch of the decoded instructions from the two threads caused by remotely injected binary code will be detected. Our design provides instruction set randomization (ISR) with minimal cost in performance, when compared with straightforward ISR implementation. The simulation results indicate that our framework incurs very small overheads and provides a protection against code injection attacks.

**Exploiting Narrow-Width Values for Improving Non-Volatile Cache Lifetime**

Guangshan Duan, Shuai Wang

Due to the high density, low leakage power consumption, and less vulnerability to soft errors, non-volatile memory technologies are among the most promising alternatives for replacing the traditional DRAM and SRAM technologies used in implementing main memory and caches in the modern microprocessor. However, one of the difficulties is the limited write endurance of most non-volatile memory technologies. In this paper, we propose to exploit the narrow-width values to improve the lifetime of non-volatile last level caches. Leading zeros masking scheme is first proposed to reduce the write stress to the upper half of the narrow-width data. To balance the write variations between the upper half and the lower half of the narrow-width data, two swap schemes, the swap on write (SW) and swap on replacement (SRep), are proposed. To further reduce the write stress to non-volatile caches, we adopt two optimization schemes, the multiple dirty bit (MDB) and read before write (RBW), to improve their lifetime. Our experimental results show that by combining all our proposed schemes, the lifetime of non-volatile caches can be improved by 245% on average.

**Partial-SET: Write Speedup of PCM Main Memory**

Bing Li, Shuchang Shan, Yu Hu, Xiaowei Li

Phase change memory (PCM) is a promising nonvolatile memory technology developed as a possible DRAM replacement. Although it offers the read latency close to that of DRAM, PCM generally suffers from the long write latency. Long write request may block the read requests on the critical path of cache/memory access, incurring adverse impact on the system performance. Besides, the write performance of PCM is very asymmetric, i.e., the SET operation (writing ‘1’) is much slower than that of the RESET operation (writing ‘0’). In this work, we re-examine the resistance transform process during the SET operation of PCM and propose a novel Partial-SET scheme to alleviate the long write latency issue of PCM. During a write access to a memory line, a short Partial-SET pulse is applied first to program the PCM cells to a pre-stable state, achieving the same write latency as RESET. The partially-SET cells are then fully programmed within the retention window to preserve the data integrity. Experimental results show that our Partial-SET scheme can improve the memory access performance of PCM by more than 45% averagely with very marginal storage overhead.

**3.6 Cyber Physical Systems: Security and Co-design**

Chair: Rolf Einst, Technische Universitat Braunschweig, DE; Anuradha Annaswamy, MIT, US

**Attack-Resilient Sensor Fusion**

Radoslav Ivanov, Miroslav Pajic, Insup Lee

This work considers the problem of attack-resilient sensor fusion in an autonomous system where multiple sensors measure the same physical variable. A malicious attacker may corrupt a subset of these sensors and send wrong measurements to the controller on their behalf, potentially compromising the safety of the system. We formalize the goals and constraints of such an attacker who also wants to avoid detection by the system. We argue that the attacker's capabilities depend on the amount of information she has about the correct sensors' measurements. In the presence of a shared bus where messages are broadcast to all components connected to the network, the attacker may consider all other measurements before sending her own in order to achieve maximal impact. Consequently, we investigate effects of communication schedules on sensor fusion performance. We provide worst- and average-case results in support of the Ascending schedule, where sensors send their measurements in a fixed succession based on their precision, starting from the most precise sensors. Finally, we provide a case study to illustrate the use of this approach.

**Bandwidth-Efficient Controller-Server Co-Design with Stability Guarantees**

Amir Aminfar, Enrico Bini, Petru Eles, Zebo Pang

Many cyber-physical systems comprise several control applications implemented on a shared platform, for which stability is a fundamental requirement. This is as opposed to the classical hard real-time systems where often the criterion is meeting the deadline. However, the stability of control applications depends not only on the delay experienced, but also on the jitter. Therefore, the notion of deadline is considered to be artificial for control applications that promotes the need for new techniques for designing cyberphysical systems. The approach in this paper is built on a server-based resource reservation mechanism, which provides compositionalty, isolation, and the opportunity of systematic controller-server co-design. We address the
controller-server co-design of such systems to obtain design solutions with the minimal bandwidth to guarantee stability.

[64] Fault-tolerant Control Synthesis and Verification of Distributed Embedded Systems
Matthias Kauer, Damos Soudabakhsh, Dip Goswami, Samrajchh Chakraborty, Anuradha M. Annaswamy
We deal with synthesis of distributed embedded control systems closed over a faulty or severely constrained communication network. Such overallocated communication networks are common in cost-sensitive domains such as automotive. Design of such systems aims to meet all deadlines following the traditional notion of schedulability. In this work, we aim to exploit robustness of the controller and propose a novel implementation approach to achieve a tighter design. Toward this, we answer two research questions: (i) given a distributed architecture, how to characterize and formally verify the bound on deadline misses, (ii) given such a bound, how to design a controller such that desired stability and Quality of Control (QoC) requirements are met. We address question (i) by modeling a distributed embedded architecture as a network of Event Count Automata (ECA), and subsequently introducing and formally verifying a property formulation with reduced complexity. We address question (ii) by introducing a novel fault-tolerant control strategy which adjusts the control input at runtime based on the occurrence of fault or drop. We show that QoC under faulty communication improves significantly using the proposed fault-tolerant strategy.

[65] Garbage Collection for Multi-version Index on Flash Memory
Kam-Yu Lam, Jian-Tao Wang, Yuan-Hao Chang, Jen-Wei Hsieh, Po-Chun Huang, Chung Keung Poon, Chun Jiang Zhu
In this paper, we study the important performance issues in using the purging-range query to reclaim old data versions to be free blocks in a flash-based multi-version database. To reduce the overheads for using the purging-range query in garbage collection, the physical block labeling (PBL) scheme is proposed to provide a better estimation on the purging version number to be used for purging old data versions. With the use of the frequency-based placement (FBP) scheme to place data versions in a block, the efficiency in garbage collection can be further enhanced by increasing the deadspaces of data versions and reducing reallocation cost especially when the sizes of the flash memory for the databases are limited.

Index Terms - Multi-version Index, Multi-version Database, Realtime Data, Flash-based Embedded Database Systems

[66] D2Cyber: A Design Automation Tool for Dependable Cybercars
Arash Munt, Farzana Koushalya
Next generation of automobiles (also known as cybercars) will increasingly incorporate electronic control units (ECUs) to implement various safety-critical functions such as x-by-wire (e.g., steer-by-wire (SBW), brake-by-wire). ISO 26262 specifies automotive safety integrity levels (ASILs) to signify the criticality associated with a function. Meeting a design's ASIL requirements at a minimum additional cost is a major challenge in cybercars design. In this paper, we propose D2Cyber - a design automation tool for cybercars that facilitates designers in selecting dependable designs by providing built-in models, easy to specify inputs, and easy to interpret outputs. D2Cyber considers the effects of temperature, electronics quality grade, and design lifetime in cybercar's design space exploration for determining a cost-effective solution and also advises on the attainable ASIL from a given design. We elaborate Markov models that form the basis of D2Cyber using SBW as a case study. We further provide evaluation insights obtained from D2Cyber.

[67] Contract-Based Design of Control Protocols for Safety-Critical Cyber-Physical Systems
Pietra Nuzzo, John B. Finn, Antonio Annapollo, Alberto L. Sangiovanni-Vincentelli
We introduce a platform-based design methodology that addresses the complexity and heterogeneity of cyber-physical systems by using assume-guarantee contracts to formalize the design process and enable control protocols in a hierarchical and compositional manner. Given the architecture of the physical plant to be controlled, the design is carried out as a sequence of refinement steps from an initial specification to a final implementation, including synthesis from requirements and mapping of higher-level functional and nonfunctional models into a set of candidate solutions built out of a library of components at the lower level. Initial top-level requirements are captured as contracts and expressed using linear temporal logic (LTL) and signal temporal logic (STL) formulas to enable requirement analysis and early detection of inconsistencies. Requirements are then refined into a controller architecture by combining reactive synthesis steps from LTL specifications with simulation-based design space exploration steps. We demonstrate our approach on the design of embedded controllers for aircraft electric power distribution.

3.7 On line Strategies for Reliability
Chairs: Fabrizio Lombardi, Northwestern University, US; Jie Han, University of Alberta, CA

[68] Spatial Pattern Prediction Based Management of Faulty Data Caches
Georgios Keramidas, Michail Mavropoulos, Anna Karvouniar, Dimitris Nikolos
Technology scaling leads to significant fault bit rates in on-chip caches. In this work, we propose a methodology to mitigate the impact of defective bits (due to permanent faults) in first-level set-associative data caches. Our technique assumes that faulty caches are enabled with the ability of disabling their defective parts at cache subblock granularity. Our experimental findings reveal that while the occurrence of hard-errors in faulty caches may have a significant impact in performance, a lot of room for improvement exists, if someone is able to take into account the spatial reuse patterns of the to-be-referenced blocks (not all the data fetched into the cache is accessed). To this end, we propose frugal PC-indexed spatial predictors (with very small storage requirements) to orchestrate the (re)placement decisions among the fully and partially unusable faulty blocks. Using cycle-accurate simulations, a wide range of scientific applications, and a plethora of cache fault maps, we showcase that our approach is able to offer significant benefits in cache performance.

[69] Combined DVFS and Mapping Exploration for Lifetime and Soft-Error Susceptibility Improvement in MPSoCs
A. Das, A. Kumar, B. Veeravalli, C. Bolchini, A. Mele
Energy and reliability optimization are two of the most critical objectives for the synthesis of multiprocessor systems-on-chip (MPSoCs). Task mapping has shown significant promise as a low cost solution in achieving these objectives as standalone or in tandem as well. This paper proposes a multiobjective design space exploration to determine the mapping of tasks of an application on a multiprocessor system and voltage/frequency level of each task (exploiting the DVFS capabilities of modern processors) such that the reliability of the platform is improved while fulfilling the energy budget and the performance constraint set by system designers. In this respect, a given MPSoC platform incorporates not only the impact of voltage and frequency on the aging of the processors (wearout effect) but also on the susceptibility to soft-errors - a joint consideration missing in all existing works in this domain. Further, the proposed exploration also incorporates soft-error tolerance by selective replication of tasks, making the proposed approach an interesting blend of reactive and proactive faulttolerance. The combined objective of minimizing core aging together with the susceptibility to transient faults under a given performance/energy budget is solved by using a multi-objective genetic algorithm exploiting tasks' mapping, DVFS and selective replication as tuning knobs. Experiments conducted with real-life and synthetic application graphs clearly demonstrate the advantage of the proposed approach.

[70] DARP: Dynamically Adaptable Resilient Pipeline Design in Microprocessors
Hu Chen, Sanghamitra Roy, Kousik Chakraborty
In this paper, we demonstrate that the sensitized path delays in various microprocessor pipe stages exhibit intriguing temporal and spatial variations during the execution of real world applications. To effectively exploit these delay variations, we propose Dynamically Adaptable Resilient Pipeline (DARP) - a series of runtime techniques to boost power performance efficiency and fault tolerance in a pipelined microprocessor. DARP employs early error prediction to avoid a major portion of the timing errors. Using a rigorous circuitarchitectural infrastructure, we demonstrate substantial improvements in the performance (9.4-20%) and energy efficiency (6.4-27.9%), compared to state-of-the-art techniques.

[71] A Fault Detection Mechanism in a Data-flow Scheduled Multithreaded Processor
Jian Fu, Qiang Yang, Raphael Poss, Chris R. Jesshope, Chunyuan Zhang
This paper designs and implements the Redundant Multi-Threading (RMT) in a Data-flow scheduled Multi-Threaded (DMT) multicores processor, called Data-flow scheduled Redundant Multi-
3.8 Hot Topic: Mission Profile Aware Design - The Solution for Successful Design of Tomorrows Automotive Electronics

Organizers: Goeran Jerke, Robert Bosch GmbH, DE; Oliver Bringmann, University of Tuebingen, DE
Chairs: Goeran Jerke, Robert Bosch GmbH, DE; Oliver Bringmann, University of Tuebingen, DE

Keywords - Electromigration, IC Design, Mission Profile, Mission Profile Aware Design, Reliability, Robustness, Validation, Verification

Mudit Bhargava, Ken Mai

Hot Topic: Multicore Systems in Safety Critical Electronic Control Units for Automotive and Avionics

Md. Tauhidur Rahman, Domenic Forte, Jim Fahrny, Mohammad Tehranipoor

Keywords - WCET, multi-core, temporal partitioning, safetycritical real-time systems

Based on a discussion of the challenges for temporal partitioning in multi-core systems, we deduce a generic architecture model. Considering the requirements for re-usability and incremental development and certification, we use this model to describe our integrated analysis approach.

Mission Profile Aware IC Design - A Case Study
Goeran Jerke, Andrew B. Kahng

Consistent consideration of mission profiles throughout a supply chain is essential for the development of robust electronic components. Consideration of mission profiles is still mainly a manual task today despite rapidly decreasing robustness margins in modern automotive semiconductor technologies. Mission profile awareness aids the automation of robustness aware design by formulating and partially automating the generation, transformation, propagation and usage of all component-specific functional loads and environmental conditions for design implementation and validation. In addition, it aids the development of electronic components in yet immature technologies or in technologies with tight parameter variation bounds. This paper introduces the general concept, requirements and context of mission profile aware design. The general design approach is presented along with key differences and enhancements to existing design approaches. A case study focusing on mission profile usage and electromigration failure avoidance is presented to demonstrate various aspects of mission profile aware design.

Keywords - PUF, reliable RO-based PUF, reliable PUF, robust PUF, aging resistant PUF, PUF reliability.

In this paper we propose to exploit so called Mission Profiles to address increasing requirements on safety and power efficiency for automotive power ICs. These Mission Profiles constrain the required device performance space to valid application scenarios. Mission Profile data can be represented in arbitrary forms like temperature histograms or cumulated drive cycle data. Hence, the derivation of realistic validation scenarios on device level requires the generation of environmental properties as e.g. temperatures, board net conditions or currents. For the assessment of real application robustness we present a methodology to extract finite state machines out of measured vehicle data and integrate them in Mission Profiles. Subsequently Markov processes are derived from these finite state machines in order to automatically generate Mission Profile compliant test scenarios for the design and verification process. As a motivating example we show industry fault cases in which missing application fitness to power transient variations finally results in device failure. Verification results based on lab data are outlined and show the benefits of a fully mission profile driven IC verification flow.

Keywords - Automotive Powers, Mission Profile, Markov Process, Finite-State Machines, Robustness, Power Transients

Application of Mission Profiles to Enable Cross-Domain Constraint-Driven Design
C. Katzsche, M.-P. Sohn, M. Olbrich, V. Meyer zu Bexten, M. Traut, E. Barke

Mission Profiles contain top-level stress information for the design of future systems. These profiles are refined and transformed to design constraints. We present methods to propagate the constraints between design domains like package and chip. We also introduce a cross-domain methodology for our corresponding constraint transformation system ConDUTC. The proposed methods are demonstrated on the basis of an automotive analog/mixed-signal application.

Index Terms - Mission Profiles and constraints, constraint transformation, cross-domain constraints, constraints in integrated circuits.

Monitoring and WCET Analysis in COTS Multi-core-Soc-Based Mixed-Criticality Systems
Jan Nowotsch, Michael Pahtschke, Ame Henrichsen, Werner Pongratz, Andreas Schacht

The performance and power efficiency of multi-core processors are attractive features for safety-critical applications, for example in avionics. But the inherent use of shared resources complicates timing analysability. In this paper we discuss a novel approach to compute the Worst-Case Execution Time (WCET) of multiple hard real-time applications scheduled on a Commercial Off-The-Shelf (COTS) multi-core processor. The analysis is closely coupled with mechanisms for temporal partitioning as, for instance, required in ARINC 653-based systems. Based on a discussion of the challenges for temporal partitioning and timing analysis in multi-core systems, we deduce a generic architecture model. Considering the requirements for re-usability and incremental development and certification, we use this model to describe our integrated analysis approach.

Keywords - WCET, multi-core, temporal partitioning, safetycritical real-time systems

Hardware Virtualization Support for Shared Resources in Mixed-Criticality Multicore Systems

Electronic architectures in modern automobiles evolve towards an hierarchical approach where functionalities from several ECUs are consolidated into few domain computers. Performance requirements directly lead to multicore solutions but also to a combination of very different requirements on such ECUs. Using virtualization in addition is one promising way of achieving segregation in time and space of shared resources. Based on examples taken from the automotive domain several concepts for efficient hardware extensions of coprocessors and I/O devices are shown in this contribution. These provide mechanisms to ensure quality of service (QoS) levels in terms of execution time, throughput and latency. The resulting infotainment architecture is a feasibility study and is integrated into a vehicle demonstrator as centralized in-vehicle platform (VCT).

Secure Device Identification
Chairs: Tim Gueneysu, RUB, DE; Guido Bertoni, TMicroelectronics, IT

ARO-PUF: An Aging-Resistant Ring Oscillator PUF Design
Md. Tauchidur Rahman, Domenic Forte, Jim Fahmy, Mohammad Tehrani,poor

Physically Uncollected Functions (PUFs) have emerged as a security block with the potential to generate chip-specific identifiers and cryptographic keys. However it has been shown that the stability of these identifiers and keys is heavily affected by aging and environmental variations. Previous techniques have mostly focused on improving PUF robustness against supply noise and temperature but aging has been largely neglected. In this paper, we propose a new aging resistant design for the popular ring-oscillator (RO)-PUF. Simulation results demonstrate that our aging resistant RO-PUF (called ARo-PUF) can produce unique, random, and more reliable keys. Only 7.7% bits get flipped on average over 10 years operation period for an ARO-PUF due to aging where the value is 32% for a conventional RO-PUF. The ARO-PUF shows an average inter-chip HD of 49.67% (close to ideal value 50%) and better than the conventional RO-PUF (~ 45%). With lower error, ARO-PUF offers ~ 24X area reduction for a 128-bit key because of reduced ECC complexity and smaller PUF footprint.

Keywords - PUF, reliable RO-based PUF, reliable PUF, robust PUF, aging resistant PUF, PUF reliability.

An Efficient Reliable PUF-Based Cryptographic Key Generator in 65nm CMOS
Mdut Bhargava, Ken Mai

Threading (DRMT). Meanwhile, it presents Asynchronous Output Comparison (AOC) for RMT techniques to avoid fault detection related inter-core communication and alleviate the performance and hardware overheads induced by output comparison. Results show that the performance overhead of DRMT is less than 60% even when the number of threads is four times the number of processing elements. Also the performance and hardware overheads of AOC are insignificant.
Physical unclonable functions (PUFs) are primitives that generate high-entropy, tamper resistant bits for use in secure systems. For applications such as cryptographic key generation, the PUF response bits must be highly reliable, consistent across multiple evaluations under voltage and temperature variations. Conventionally, error correcting codes (ECC) have been used to improve response reliability, but these techniques have significant area, power, and delay overheads and are vulnerable to information leakage. In this work, we present a highly reliable, PUF-based, cryptographic key generator that uses no ECC, but instead uses built-in self-test to determine which PUF bits are reliable and only uses those bits for key generation. We implemented a prototype of the key generator in a 65nm bulk CMOS testchip. The key generator generates 1213 bits in an area of <50µm² with a measured bit error rate of < 5 x 10⁻⁶ in both the nominal and worst case corners (100k measurements each). This is equivalent to a 128-bit key failure rate of < 10⁻⁶. The system can generate a 128-bit key in 1.15µs. Finally, we present a realization of a "strong" PUF that uses 128 of these highly reliable bits in conjunction with an Advanced Encryption Standard (AES) cryptographic primitive and has a response time of 40ns and is realized in an area of 84kµm².

4.5 Memory System Architectures

Chairs: Muhammad Shafique, Karlsruhe Institute of Technology, DE; Cristina Silva, Politecnico di Milano, IT

(85) Achieving Efficient Packet-based Memory System by Exploiting Correlation of Memory Requests
Tianyue Lu, Licheng Chen, Mingyu Chen

Packet-based interface is a trend for future memory system to alleviate memory capacity and bandwidth bottlenecks. On the other hand fine-grained memory access has been proven to efficiently reduce memory power. However leveraging both these two technologies will result in high packet overhead, because previous implementations of packet-based interface all adopt a simple design that a single packet is dedicated to a single request (SPSR). In this paper, we propose three optimizations to overcome the problem by exploiting correlations of memory requests. First, we propose a novel single packet multiple requests (SPMR) interface that encapsulates multiple requests into a single packet header and tail. Second, we
propose an adaptive address compression mechanism within a packet by adopting a base-difference algorithm. Third, we propose a mechanism to merge multiple memory requests with continuous access addresses into a single request before packing. By this way, the granularity constraint of cache line size is broken to enable efficiently row buffer scheduling. The experimental results show that, for certain memory-intensive workloads, the optimizations can effectively reduce packet overhead by about 53.9% and improve system performance by about 63.6% in average.

[88] ALLARM: Optimizing Sparse Directories for Thread-Local Data
Amitabraj Roy, Timothy M. Jones
Large-scale cache-coherent systems often impose unnecessary overhead on data that is thread-private for the whole of its lifetime. These include resources devoted to tracking the coherence state of the data, as well as unnecessary coherence messages sent out over the interconnect. In this paper we show how the memory allocation strategy for non-uniform memory access (NUMA) systems can be exploited to remove any coherencerelated traffic for thread-local data, as well removing the need to track those cache lines in sparse directories. Our strategy is to allocate directory state only on a miss from a node in a different affinity domain from the directory. We call this ALLocate on Remote Miss, or ALLARM. Our solution is entirely backward compatible with existing operating systems and hardware, and provides a means to obsolete cache coherence into the many-core era. On a mix of SPLASH2 and PARSEC workloads, ALLARM is able to improve performance by 13% on average while reducing dynamic energy consumption by 9% in the on-chip network and 15% in the directory controller. This is achieved through a 46% reduction in the number of sparse directory entries evicted.

[89] Introducing Thread Critically Awareness in Prefetcher Aggressiveness Control
Biswabandha Panda, Shankar Balachandran
A single parallel application running on a multicore system shows sub-linear speedup because of slow progress of one or more threads known as critical threads. Some of the reasons for the slow progress of threads are (1) load imbalance, (2) frequent cache misses and (3) effect of synchronization primitives. Identifying critical threads and minimizing their cache miss latencies can improve the overall execution time of a program. One way to hide and tolerate the cache misses is through hardware prefetching. Hardware prefetching is one of the most commonly used memory latency hiding techniques. Previous studies have shown the effectiveness of hardware prefetchers for multithreaded applications (multiple sequential applications running independently on different cores). In contrast to multithreaded workloads, the performance of a single parallel application depends on the progress of slow progress (critical) threads. This paper introduces a prefetcher aggressiveness control mechanism called Thread Critically-aware Prefetcher Aggressiveness Control (TCPAC). TCPAC controls the aggressiveness of the prefetchers at the L2 prefetching controllers (known as TCPAC-P), DRAM controller (known as TCPAC-D) and at the Last Level Cache (LLC) controller (known as TCPAC-C) using prefetch accuracy and thread progress. Each TCPAC sub technique outperforms the respective state-of-the-art techniques such as HPAC [2], PADC [4] and PACMan [3] and the combination of all the TCPAC sub-techniques namely TCPAC-PDC outperforms the combination of HPAC, PADC, and PACMan. On an average, on a 8 core system, in terms of improvement in execution time, TCPAC-PDC outperforms the combination of HPAC, PADC, and PACMan by 7.61%. For 12 and 16 cores TCPAC-PDC beats the state-of-the-art combinations by 7.21% and 8.32% respectively.

[88] A Multi Banked - Multi Ported - non Blocking Shared L2 Cache for MPSoC Platforms
Igor Loi, Luca Benini
On-chip L2 cache architectures, well established in high-performance parallel computing systems, are now becoming a performance-critical component also for multi/many-core architectures targeted at low-power, embedded applications. The very stringent requirements on power and cost of these systems result in one of the key challenges in many-core designs, mandating the deployment of highly efficient L2 caches. In this perspective, sharing the L2 cache layer among all system cores has important advantages, such as increased utilization, fast inter-core communication, and reduced aggregate footprint because no undesired replication of lines occurs. This paper presents a novel architecture for a shared L2 cache system with multi-port and multi-bank features. We target this L2 cache to a many-core platform based on hierarchical cluster structure that does not employ private data caches, and therefore does not require complex coherence mechanisms. In fact, our shared L2 cache can be seen logically as a Last Level Cache (LLC) adopting the terminology of high-performance many-core products, although in these latter the LLC is often an L3 layer. Our experimental results show a maximum aggregate bandwidth of 28GB/s (89% of the maximum channel capacity) for 100% hit traffic with random-banking conflicts, as a realistic case. Physical implementation results in 28nm Fully-Depleted-Silicon-on-Insulator (FDSOI) show that our L2 cache can operate at up to 1GHz with a memory density loss of only 20% with respect to an L2 scratchpad for a 2 MB configuration.

[90] DRAM-based Coherent Caches and How to Take Advantage of the Coherence Protocol to Reduce the Refresh Energy
Ziron Laksh, Raghunath Gaital
Recent technology trends has turned DRAMs into an interesting candidate to substitute traditional SRAM-based on-chip memory structures (i.e., register file, cache memories). Nevertheless, a major problem to introduce these cells is that they lose their state (i.e., value) over time, and they have to be refreshed. This paper proposes the implementation of coherent caches with DRAM cells. Furthermore, we propose to use the coherence state to tune the refresh overhead. According to our analysis, an average of up to 57% of refresh energy can be saved. Also, comparing to the caches implemented in SRAMs total energy savings are on average up to 39% depending of the refresh policy with a performance loss below 8%.

Index Terms - FinFETs, 6T SRAM, 3T DRAM, retention time, cache coherence

[91] Reducing Set-Associative L1 Data Cache Energy by Early Load Data Dependency Detection (ELD)
Alen Bardizbanyan, Magnus Själander, David Whalley, Per Larsson-Edefors
Fast set-associative level-one data caches (L1 Dcs) access all ways in parallel during load operations for reduced access latency. This is required in order to resolve data dependencies as early as possible in the pipeline, which otherwise would suffer from stall cycles. A significant amount of energy is wasted due to this fast access, since the data can only reside in one of the ways. While it is possible to reduce L1 DC energy usage by accessing the tag and data memories sequentially, hence activating only one data way on a tag match, this approach significantly increases execution time due to an increased number of stall cycles. We propose an early load data dependency detection (ELD) technique for in-order pipelines. This technique makes it possible to detect if a load instruction has a data dependency with a subsequent instruction. If there is no such dependency, then the tag and data accesses for the load are sequentially performed so that only the data way in which the data resides is accessed. If there is a dependency, then the tag and data arrays are accessed in parallel to avoid introducing additional stall cycles. For the MBench benchmark suite, the ELD3 technique enables about 49% of all load operations to access the L1 DC sequentially. Based on 65-nm data using commercial SRAM blocks, the proposed technique reduces L1 DC energy by 13%.

[92] Distributed Cooperative Shared Last-Level Caching in Tiled Multiprocessor System on Chip
Preethi P. Damodaran, Stefan Wallentowitz, Andreas Herkersdorf
In a shared-memory based tiled many-core system-on-chip architecture, memory accesses present a huge performance bottleneck in terms of access latency as well as bandwidth requirements. The best practice approach to address this issue is to provide a multi-level cache hierarchy and a suitable cache-coherency mechanism. This paper presents a method to increase the memory access performance in distributed-coherence-coherency-protocol based tiled many-core systems. The proposed method introduces an alternate design for the system-wide shared last-level caches (LLC) placed between the memory and the node private caches (NPC). The proposed system-wide shared LLC layer is distributed over the entire network and it interacts with the home directories of specific cache lines. Results from simulating SPEC2000 benchmark applications executed on a SystemC model of the proposed design show a minimum performance improvement of 20-25% when compared to a model without the shared cache layer at the expense of an additional 2% of the total cache memory space (NPC + LLC memory). In addition, the proposed design shows a minimum 7-15% and an average 14-15% improvement in performance in comparison to centralized system-wide shared LLC of equivalent size and dynamic mapped distributed LLC of equivalent size respectively.

4.6 Code Generation and Optimization for Embedded Platforms
Chairs: Heiko Falk, Ulm University, DE; Florence Maraninchi, Grenoble IMP/VERIMAG, FR

[92] EATBit: Effective Automated Test for Binary Translation with High Code Coverage


Dependable System Design

Trojans. To this end, we have designed a Trojan using power-gating techniques and demonstrate that it can be masked from advanced side-channel fingerprinting detection while fabricated integrated circuits. However, these post-fabrication trust evaluation methods overlook the capabilities of advanced design skills that attackers can use in designing sophisticated

Yier Jin, Dean Sullivan

spills by 12.5%, 11.7%, 12.7%, for three different processor models, respectively.

linear scan register allocator without considering lifetime holes by using the MediaBench II benchmark suite. Our simulation results show that our register allocator reduces the number of

Xuemeng Zhang, Hui Wu, Haiyan Sun, Jingling Xue

The experimental results show that the proposed method can significantly reduce the write variation on PCM cells and enhance the lifetime of PCM memory.

good scalability. However, PCM's low endurance constrains its practical applications. In this paper, we propose a wear leveling aware dynamic stack to extend PCM's lifetime when it is

Qingan Li, Yanxiang He, Yong Chen, Chun Jason Xue, Nan Jiang, Chao Xu

(up to 32%) energy savings with the proposed optimization strategy.

determining the placement of wakelock statements corresponding to the uses of devices in an application. Our experimental evaluation on a set of Android applications show significant

The Android OS Power Manager provides programming interface routines called wakelocks for controlling the activation state of devices on a mobile system. An appropriate placement of

wakelock acquire and release functions in the application can make a significant difference to the energy consumption. In this paper, we propose a data flow analysis based strategy for

The Android OS Power Manager provides programming interface routines called wakelocks for controlling the activation state of devices on a mobile system. An appropriate placement of

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

Richard Membarth, Oliver Reiche, Frank Harrig, Jürgen Teich

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

The Android OS Power Manager provides programming interface routines called wakelocks for controlling the activation state of devices on a mobile system. An appropriate placement of

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of different target platforms. However, this comes at

Design of Safety Critical Systems by Refinement

Alex Illasov, Arseniy Akhseyev, Davide Scolov, Andrey Mikhov

An increasingly large number of safety-critical embedded systems rely on software to prevent and mitigate hazards occurring due to design errors and unexpected interactions of the system with its users and the environment. Implementing a safety instrumented function in the way advocated by the traditional software methods requires an intimate understanding and thorough validation of a complex ecosystem of programming languages, compilers, operating systems and hardware. We propose to consider an alternative where a system designer, for each individual problem, creates in a correct-by-construction manner both the design of a system and its compilation and execution infrastructure. This permits an uninterrupted chain of a formal correctness argument spanning from formalised requirements all the way to the gate-level characterisation of an execution environment. The past decade of advances in verification technology turned the mechanical verification of large-scale models into a reality while the pressure of certification makes the cost of a formally verified development route increasingly acceptable. The proposal fits the Grand Challenge for Computer Research posed by Hoare in 2003, namely, development of a Verifying Compiler which not only mechanically translates a given program from one language to another but also verifies its correctness according to a formal specification. This allows meeting the most stringent software certification requirements such as SIL 4. We illustrate the vision with a small case-study developed using the Event-B modelling notation and tools.

Energy Optimization in Android Applications through WakeLock Placement

Faisal Alam, Preeti Ranjan Panda, Nithi Tripathi, Namita Sharma, Sanjay Narayan

Energy efficiency is a critical factor in mobile systems, and a significant body of recent research efforts has focused on reducing the energy dissipation in mobile hardware and applications. The Android OS Power Manager provides programming interface routines called wakelocks for controlling the activation state of devices on a mobile system. An appropriate placement of wakelock acquire and release functions in the application can make a significant difference to the energy consumption. In this paper, we propose a data flow analysis based strategy for determining the placement of wakelock statements corresponding to the uses of devices in an application. Our experimental evaluation on a set of Android applications show significant (up to 32%) energy savings with the proposed optimization strategy.

A Wear-Leveling-Aware Dynamic Stack for PCM Memory in Embedded Systems

Qingan Li, Yanxiang He, Yong Chen, Chun Jason Xue, Nan Jiang, Chao Xu

Phase Change Memory (PCM) is a promising DRAM replacement in embedded systems due to its attractive characteristics such as extremely low leakage power, high storage density and good scalability. However, PCM's low endurance constrains its practical applications. In this paper, we propose a wear leveling aware dynamic stack to extend PCM's lifetime when it is adopted in embedded systems as main memory. Through a dynamic stack, the memory space is circularly allocated to stack frames, and thus an even usage of PCM memory is achieved. The experimental results show that the proposed method can significantly reduce the write variation on PCM cells and enhance the lifetime of PCM memory.

Lifetime Holes Aware Register Allocation for Clustered VLIW Processors

Xuemeng Zhang, Hui Wu, Haiyan Sun, Jingling Xue

This paper presents an on-the-fly register allocator which dynamically detects and utilizes lifetime holes for clustered VLIW processors. A lifetime hole is an interval in which a variable does not contain a valid value. A register holding a lifetime hole can be allocated to another variable whose live range fits in the lifetime hole, leading to more efficient utilisation of registers. We propose efficient techniques for dynamically utilizing lifetime holes and incorporate these techniques into our on-the-fly register allocator. We have simulated our register allocator and a linear scan register allocator without considering lifetime holes by using the MediaBench II benchmark suite. Our simulation results show that our register allocator reduces the number of spills by 12.5%, 11.7%, 12.7%, for three different processor models, respectively.

Index Terms - register allocation; lifetime hole; live range; clustered VLIW processor; inter-cluster communication

4.7 Dependable System Design

Chair: Yiorgos Makris, University of Texas at Dallas, US; Michael Nicolaides, TIMA, FR

Real-Time Trust Evaluation in Integrated Circuits

Yier Jin, Dean Sullivan

The use of side-channel measurements and fingerprinting, in conjunction with statistical analysis, has proven to be the most effective method for accurately detecting hardware Trojans in fabricated integrated circuits. However, these post-fabrication trust evaluation methods overlook the capabilities of advanced design skills that attackers can use in designing sophisticated Trojans. To this end, we have designed a Trojan using power-gating techniques and demonstrate that it can be masked from advanced side-channel fingerprinting detection while
downtime. We then propose a real-time trust evaluation framework that continuously monitors the on-board global power consumption to monitor chip trustworthiness. The measurements obtained corroborate our framework effectiveness for detecting Trojans. Finally, the results presented are experimentally verified by performing measurements on fabricated Trojan-free and Trojan-infected variants of a reconfigurable linear feedback shift register (LFSR) array.

Verification-guided Voter Minimization in Triple-Modular Redundant Circuits

Dmytro Buryaev, Pascal Fadet, Alan Girault

We present a formal approach to minimize the number of voters in triple-modular redundant sequential circuits. Our technique actually works on a single copy of the circuit and considers a user-defined fault model (under the form "at most 1 bit-flip every k clock cycles"). Verification-based voter minimization guarantees that the resulting circuit (i) is fault tolerant to the soft-errors defined by the fault model and (ii) is functionally equivalent to the initial one. Our approach operates at the logic level and takes into account the input and output interface specifications of the circuit. It implements makes use of graph traversal algorithms, fixed-point iterations, and BDDs. Experimental results on the ITC'99 benchmark suite indicate that our method significantly decreases the number of inserted voters which entail a hardware reduction of up to 55% and a clock frequency increase of up to 35% compared to full TMV. We address scalability issues arising from formal verification with approximations and assess their efficiency and precision.


Yue Gao, Sandeep K. Gupta, Yanzhi Wang, Massoud Pedram

For modern high performance systems, aggressive technology and voltage scaling has drastically increased their susceptibility to soft errors. At the grand scale of cloud computing, it is clear that soft error induced failures will occur far more frequently, but it is unclear as to how to effectively apply current error detection and fault tolerance techniques in scale. In this paper, we focus on energy-aware fault tolerant scheduling in public, multi-user cloud systems, and explore the three-way tradeoff between reliability (in terms of soft error resiliency), performance and energy. Through a systematically optimized resource allocation, error detection approach selection, virtual machine placement, spatial/temporal redundancy augmentation and task scheduling process, the cloud service provider can achieve high error coverage and fault tolerance confidence while minimizing global energy costs under user deadline constraints. Our scheduling algorithm includes a static scheduling phase that operates on task graph based workload inputs prior to execution, and a light-weight dynamic scheduler that migrates tasks during execution in case of excessive re-executions. All schedules are evaluated on a runtime simulation engine that (1) mimics the performance fluctuations in cloud systems, and (2) supports the injection of arbitrary fault patterns. Compared to current virtual machine or task replication techniques, we are able to reduce overall application failure rates by over 50% with approximately 76% total energy overhead.

A Low-Power, High-Performance Approximate Multiplier with Configurable Partial Error Recovery

Cong Liu, Jie Han, Fabrizio Lombardi

Approximate circuits have been considered for error-tolerant applications that can tolerate some loss of accuracy with improved performance and energy efficiency. Multipliers are key arithmetic units in applications such as digital signal processing (DSP). In this paper, a novel approximate multiplier is proposed for high-performance DSP applications. This multiplier leverages a newly-designed approximate adder that limits the carry propagation to the nearest neighbors for fast partial product accumulation. Different levels of accuracy can be achieved through a configurable error recovery by using different numbers of most significant bits (MSBs) for error reduction. The approximate multiplier has a low mean error distance, i.e., most of the errors are not significant in magnitude. Compared to the Wallace multiplier, a 16-bit approximate multiplier implemented in a 28nm CMOS process shows a reduction in delay and power of 20% and up to 69%, respectively. It is shown that by utilizing an appropriate error recovery, the proposed approximate multiplier achieves similar processing accuracy as traditional exact multipliers but with significant improvements in power and performance.

5.1 Hot Topic: Predictable Multi-Core Computing

Organizer: Jürgen Teich, Erlangen-Nuremberg U, DE
Chairs: Petru Eles, Linköping U, SE; Jürgen Teich, Erlangen-Nuremberg U, DE

Impact of Resource Sharing on Performance and Performance Prediction

Jan Reineke, Reinhard Wilhelm

Multi-core processors are increasingly considered as execution platforms for embedded systems because of their good performance/energy ratio. Many applications implemented on multi-core platforms are safety- and some also time-critical. A critical issue for these applications is the reduced predictability of such systems resulting from the interference of different applications on shared resources. These interferences can be at least of two kinds: Several applications may request a resource at the same time, but the resource can only admit one access at a time. As a consequence, an arbitration mechanism may delay the request of all but one application, thus slowing down the other applications. This is the case of resources like buses, typically called bandwidth resources. On the other hand, one application may also change the state of a shared resource such that another application using that resource will suffer from a slowdown. This is the case with shared memories, such as shared caches and shared dynamic random-access memories, which fall into the class of storage resources. Interference on shared resources makes worst-case execution time (WCET) analysis of applications more difficult since a task or a thread can no longer be analyzed for its timing behavior independently of interferences. All potential interferences slowing down (or speeding up) the task under analysis have to be considered. This leads to a combinatorial explosion of the analysis complexity, as all possible interleavings of different threads have to be analyzed.

The survey [1] considers several aspects of the execution of sets of tasks on multi-core platforms that have to do with the interference of the tasks on shared resources. One question is how the actual performance of tasks is slowed down by other co-running tasks. Another is how to compute bounds on the slowdown in order to derive sound guarantees for the timing behavior. A major problem is the increased complexity of this task compared to the single-task single-core case. This has led to the situation that industry is developing embedded systems for multi-core platforms while there exist no timing analysis methods and tools that are both sound and precise.

Time-Critical Computing on a Single Chip Massively Parallel Processor

Benoît Dupont de Dinechin, Duco van Amstel, Marc Pouliès, Guillaume Lager

The requirement of high performance computing at low power can be met by the parallel execution of an application on a possibly large number of programmable cores. However, the lack of accurate timing properties may prevent parallel execution from being applicable to time-critical applications. We illustrate how this problem has been addressed by suitably designing the architecture, implementation, and programming model of the Kaisys MPPA®-256 single-chip many-core processor. The MPPA®-256 (Multi-Purpose Processing Array) processor integrates 256 processing engines (PE) cores and 32 resource management (RM) cores on a single 28nm CMOS chip. These VLIW cores are distributed across 16 compute clusters and 4 I/O subsystems, each with a locally shared memory. On-chip communication and synchronization are supported by an explicitly addressed dual network-on-chip (NoC), with one node per compute cluster and 4 nodes per I/O subsystem. Off-chip interfaces include DDR, PCI and Ethernet, and a direct access to the NoC for low-latency processing of data streams. The key architectural features that support time-critical applications are timing compositional cores, independent memory banks inside the compute clusters, and the data NoC whose guaranteed services are determined by network calculus. The programming model provides communicators that effectively support distributed computing primitives such as remote writes, barrier synchronizations, active messages, and communication by sampling. POSIX time functions expose synchronous clocks inside compute clusters and mesosynchronous clocks across the
5.2 Reliable Systems in the Age of Variability

Keywords - Variability; timing errors; error recovery; temporal memoization; value locality; GPGPUs

This paper proposes a design-time (offline) analysis technique to determine application task mapping and scheduling on a multiprocessor system and the voltage and frequency levels of all cores (offline DVFS) that minimize application computation and communication energy, simultaneously minimizing processor aging. The proposed technique incorporates (1) the effect of the voltage and frequency on the temperature of a core; (2) the effect of neighboring cores’ voltage and frequency on the temperature (spatial effect); (3) pipelined execution and cyclic dependencies among tasks; and (4) the communication energy component which often constitutes a significant fraction of the total energy for multimedia applications. The temperature model proposed here can be easily integrated in the design space exploration for multiprocessor systems. Experiments conducted with MPEG-4 decoder on a real system demonstrate that the overall optimization technique achieves 40% savings in energy consumption with 6% increase in system lifetime.

Keywords - Variability; timing errors; error recovery; temporal memoization; value locality; GPGPU

5.3 Systematic Design of Resilient ICs

Keywords - Site-channel attacks, Hardware Trojans, fault injection, counterfeiting.

This paper discusses how to combine this policy with an optimization method for the partitioning of tasks to cores as well as the static mapping of memory blocks, i.e., task data and communication buffers, to the banks of a shared memory architecture. Optimization is performed at design time targeting at minimizing the worst-case response times of tasks and achieving efficient resource utilization. The proposed optimization method is evaluated using an industrial application.

5.4 Reconfigurable Arrays for Temporal Fault Tolerance

Keywords - Site-channel attacks, Hardware Trojans, fault injection, counterfeiting.

This paper presents a novel class of exceptions called Reliability-Aware Exceptions (RAEs) which provide the ability to distinguish intermittent faults in microprocessor array structures. The RAE handlers have the ability to manipulate microprocessor array structures to recover from all three categories of faults. Using RAEs, we demonstrate that the reliability of two representative microarchitecture structures, load/store queue and reorder buffer in an out-of-order processor, is improved by average factors of 1.3 and 1.95, respectively.

Keywords - Reliability, array structure, fault injection, de-configuration

5.5 Mixing Mixed-Criticality Applications on Multi-Core Architectures

Keywords - Mixed-Criticality, scheduling, multi-core, data placement, performance evaluation

A common trend in real-time embedded systems is to integrate multiple applications on a single platform. Such systems are known as mixed-criticality (MC) systems when the applications are characterized by different criticality levels. Nowadays, multicore platforms are promoted due to cost and performance benefits. However, certification of multicoresystems is challenging as concurrently executed applications of different criticalities may block each other when accessing shared platform resources. Most of the existing research on multicorescheduling ignores the efforts of resource sharing on the response times of applications. Recently, a MC scheduling strategy was proposed, which explicitly accounts for these effects. This paper discusses how to combine this policy with an optimization method for the partitioning of tasks to cores as well as the static mapping of memory blocks, i.e., task data and communication buffers, to the banks of a shared memory architecture. Optimization is performed at design time targeting at minimizing the worst-case response times of tasks and achieving efficient resource utilization. The proposed optimization method is evaluated using an industrial application.
Reliability is a major concern in multiprocessors. Dynamic Reliability Management (DRM) aims at trading off processor performance with lifetime. The state-of-the-art publications study only the theory supported by simulation. This paper presents the first complete software implementation, working on a real hardware, of a low-overhead, Android-compatible workload aware DRM Governor for mobile multiprocessors. We discuss the design challenges and the run-time overhead involved. We show the effectiveness of our governor in guaranteeing the predefined target lifetime and show that it achieves up to 100% of lifetime improvement with respect to traditional governors, while providing comparable performance for critical applications.

Yield and Timing Constrained Spare TSV Assignment for Three-Dimensional Integrated Circuits
Yu-Guang Chen, Kuan-Yu Lai, Ming-Chao Lee, Yi Yu Shi, Ming-Kai Hon, Shih-Chieh Chang

5.4 Prediction and Optimization of Timing Variations

Chairs: Antonio Rubio, UPC Barcelona, ES; Mansa Lopez Valles, UPM Madrid, ES

Efficient High-Sigma Yield Analysis for High Dimensional Problems
Moning Zhang, Zuochang Ye, Yan Wang

Compiler-Driven Dynamic Reliability Management for On-Chip Systems under Variabilities
Sameen Rehman, Florian Kriebel, Muhammad Sharifique, Jörg Henkel

Sub-threshold Logic Circuit Design Using Feedback Equalization
Mahmoud Zangeneh, Ajay Ashi

Stochastic Analysis of Bubble Razor
Guowei Zhang, Peter Beerel

Minimizing State-of-Health Degradation in Hybrid Electrical Energy Storage Systems with Arbitrary Source and Load Profiles
Yanzhi Wang, Xue Lin, Qiong Xie, Naehyuck Chang, Massoud Pedram

Dynamic Flip-Flop Conversion to Tolerate Process Variation in Low Power Circuits
Mehrdad Nejat, Bijan Akhazadeh, Ali Afzal Kusha

A novel time borrowing method called dynamic Flip-Flop conversion is presented in this paper. A timing violation predictor detects the violations halfway in the critical path and dynamically converts the critical Flip-Flop to a latch. This way, time borrowing benefits of latches are utilized in a Flip-Flop based design which is more adaptable with Computer-Aided-Design tools. The overhead of this method is smaller than that of similar methods due to the elimination of delay elements. According to the post-synthesis simulations and Monte-Carlo analysis of Spice simulations on some ITC '99 benchmark circuits, the power overhead of the proposed method is about 15% and 19% smaller than that of Soft-Edge-Flip-Flop and Dynamic-Clock-Stretching circuits respectively in a simple case of about 40% yield improvement. This overhead would be relatively even smaller for higher performance and yield improvements.

Keywords - Timing violation, Flip-Flop, Latch, Timing violation
5.5 Boosting the Scalability of Formal Verification Technologies

Chairs: Fahim Rahim, Atrenta, FR; Bernd Becker, University of Freiburg, DE

Sebastiaan J.C. Joosten, Julien Schmitz

In the realm of multi-core processors and systems-on-chip, communication fabrics constitute a key element. A large number of queues and distributed control are two important aspects of this class of designs. These aspects make decomposition and abstraction techniques difficult to apply. For this class of designs, the application of formal methods is a real challenge. In particular, the verification of liveliness properties is often intractable. Communication fabrics can be seen as a set of queues and flops interconnected by combinatorial logic. Based on this simple but powerful observation, we propose a novel method for liveliness verification. Our method exploits the notion of Transfer Level designs. The essential aspects of our approach are (1) to abstract away from the details of queue implementations and (2) an efficient encoding of liveliness properties in an SMT instance. Experimental results are promising. Designs with hundreds of queues can be analysed for liveliness within minutes.

Tobias Welp, Andreas Kuehmann

We present a novel, sound, and complete algorithm for deciding safety properties in programs with static memory allocation. The new algorithm extends the program verification paradigm using loop invariants presented in [1] with a counterexample guided abstraction refinement (CEGAR) loop [2] where the refinement is achieved by strengthening loop invariants using the QF BV generalization of Property Directed Reachability (PDR) discussed in [3, 4]. We compare the algorithm with other approaches to program verification and report experimental results.

Christoph Scholl, Florian Pigorsch, Stefan Dech, Ernst Althaus

Craig interpolation has turned out to be an essential method for many applications in formal verification. In this paper, we focus on the computation of simple interpolants for the theory of linear arithmetic with rational coefficients. We successfully minimize the number of linear constraints in the final interpolant by several methods including proof transformations, linear programming and SMT solving. Experimental results comparing the approach to standard methods from the literature prove the effectiveness of the approach and show reductions of up to 70% in the number of linear constraints.

G. Cabodi, P. Pasini, S. Quer, D. Vendraamnettu

In the framework of symbolic model checking, BDD-based approximate reachability is potentially much more scalable than its exact counterpart. However, its practical applicability is highly limited by its static approach to abstraction, and the intrinsic difficulty to find an acceptable trade-off between accuracy and memory/time complexity. In this paper, we apply SAT-based cube generalization, a core step of the ICI model checking algorithm, to BDD-based overapproximate reachability analysis. More specifically, we use cube generalization, in both its inductive and non-inductive versions, to tighten BDD-based overapproximate representations of state sets computed by Machine by Machine (MBM) and Frame by Frame (FFB) algorithms. Experimental results confirm that this approach can provide tighter representations of reachable state sets and more powerful fully BDD-based engines, as well as potential applications of BDDs as invariants or constraints in SAT-based model checking.

Mikam Leeaer, Saami Mukheerje, Jaideep Ramachandran, Thomas Wahl

Floating-point arithmetic is widely used in scientific computing. While many programmers are subliminally aware that floating-point numbers only approximate the reals, few are cognizant of the dangers this entails for programming. Such dangers range from tolerable rounding errors in sequential programs, to unexpected, divergent control flow in parallel code. To address these problems, we present a decision procedure for floating-point arithmetic (FPA) that exploits the proximity to real arithmetic (RA), via a loss-less reduction from FPA to RA. Our procedure does not involve any form of bit-blasting or bit-vectorization, and can thus generate much smaller back-end decision problems, albeit in a more complex logic. This tradeoff is beneficial for the exact and reliable analysis of parallel scientific software, which tends to give rise to large but benignly structured formulas. We have implemented a prototype decision engine and present encouraging results analyzing such software for numerical accuracy.

5.6 Emerging Logic Technologies

Chairs: Mehdi Tahoori, KIT, DE; Marco Ottavi, University of Rome "Tor Vergata", IT

Mohammad D. Motlagh, Arjun Rappallai, Chris Dewey

Resonance energy transfer (RET) circuits are networks of photo-active molecules that can implement arbitrary logic functions. The nanoscale size of these structures can bring high-density computation to new domains, e.g., in vivo sensing and computation. A key challenge in the design of a RET network is to find, among a huge set of configurations (i.e., design space), the optimum choice and arrangement of molecules on a nanostucture. The prohibitively large size of the design space makes it impractical to evaluate every possible configuration, motivating the need for design-space pruning to be integrated into the design flow. To this end, we have developed a computer-aided design framework, called RETLab, that enables structured pruning of the design space to extract a sufficiently small subset, which is fully evaluated and ranked based on user-defined metrics to yield the best configuration. More importantly, we have developed a new RET-simulation algorithm which is several orders of magnitude (e.g., for a 4-node network, one million times) faster than the conventional Monte-Carlo-based simulation (MCS). This speedup in configuration evaluation enables a significantly more extensive design-space exploration with fewer and less constrained heuristics, compared to existing RET network design methods which are ad-hoc and rely on MCS for configuration evaluation. Designs of hundreds of queues can be analysed for liveliness within minutes.

Tobias Welp, Andreas Kuehmann

We present a novel, sound, and complete algorithm for deciding safety properties in programs with static memory allocation. The new algorithm extends the program verification paradigm using loop invariants presented in [1] with a counterexample guided abstraction refinement (CEGAR) loop [2] where the refinement is achieved by strengthening loop invariants using the QF BV generalization of Property Directed Reachability (PDR) discussed in [3, 4]. We compare the algorithm with other approaches to program verification and report experimental results.

Christoph Scholl, Florian Pigorsch, Stefan Dech, Ernst Althaus

Craig interpolation has turned out to be an essential method for many applications in formal verification. In this paper, we focus on the computation of simple interpolants for the theory of linear arithmetic with rational coefficients. We successfully minimize the number of linear constraints in the final interpolant by several methods including proof transformations, linear programming and SMT solving. Experimental results comparing the approach to standard methods from the literature prove the effectiveness of the approach and show reductions of up to 70% in the number of linear constraints.

G. Cabodi, P. Pasini, S. Quer, D. Vendraamnettu

In the framework of symbolic model checking, BDD-based approximate reachability is potentially much more scalable than its exact counterpart. However, its practical applicability is highly limited by its static approach to abstraction, and the intrinsic difficulty to find an acceptable trade-off between accuracy and memory/time complexity. In this paper, we apply SAT-based cube generalization, a core step of the ICI model checking algorithm, to BDD-based overapproximate reachability analysis. More specifically, we use cube generalization, in both its inductive and non-inductive versions, to tighten BDD-based overapproximate representations of state sets computed by Machine by Machine (MBM) and Frame by Frame (FFB) algorithms. Experimental results confirm that this approach can provide tighter representations of reachable state sets and more powerful fully BDD-based engines, as well as potential applications of BDDs as invariants or constraints in SAT-based model checking.

Mikam Leeaer, Saami Mukheerje, Jaideep Ramachandran, Thomas Wahl

Floating-point arithmetic is widely used in scientific computing. While many programmers are subliminally aware that floating-point numbers only approximate the reals, few are cognizant of the dangers this entails for programming. Such dangers range from tolerable rounding errors in sequential programs, to unexpected, divergent control flow in parallel code. To address these problems, we present a decision procedure for floating-point arithmetic (FPA) that exploits the proximity to real arithmetic (RA), via a loss-less reduction from FPA to RA. Our procedure does not involve any form of bit-blasting or bit-vectorization, and can thus generate much smaller back-end decision problems, albeit in a more complex logic. This tradeoff is beneficial for the exact and reliable analysis of parallel scientific software, which tends to give rise to large but benignly structured formulas. We have implemented a prototype decision engine and present encouraging results analyzing such software for numerical accuracy.

Mohammad D. Motlagh, Arjun Rappallai, Chris Dewey

Resonance energy transfer (RET) circuits are networks of photo-active molecules that can implement arbitrary logic functions. The nanoscale size of these structures can bring high-density computation to new domains, e.g., in vivo sensing and computation. A key challenge in the design of a RET network is to find, among a huge set of configurations (i.e., design space), the optimum choice and arrangement of molecules on a nanostucture. The prohibitively large size of the design space makes it impractical to evaluate every possible configuration, motivating the need for design-space pruning to be integrated into the design flow. To this end, we have developed a computer-aided design framework, called RETLab, that enables structured pruning of the design space to extract a sufficiently small subset, which is fully evaluated and ranked based on user-defined metrics to yield the best configuration. More importantly, we have developed a new RET-simulation algorithm which is several orders of magnitude (e.g., for a 4-node network, one million times) faster than the conventional Monte-Carlo-based simulation (MCS). This speedup in configuration evaluation enables a significantly more extensive design-space exploration with fewer and less constrained heuristics, compared to existing RET network design methods which are ad-hoc and rely on MCS for configuration evaluation. Designs of hundreds of queues can be analysed for liveliness within minutes.
devices with perpendicular magnetic anisotropy (PMA) (i.e., out-of-plane magnetization states) and discuss their behavior when utilized in 3D designs. Furthermore, we provide a systematic design approach for 3D NML circuits using a threshold full adder as a case study. We compare our 3D adder to 2D adders to highlight the benefits of 3D NML circuits, which include simpler signal routing and a smaller area footprint.

In this paper, we present a highly accurate dosedilatate compact model for Schottky-barrier-type Graphene Nano-Ribbon Field-Effect Transistors (SB-GNRFETs). This is a physics-based analytical model for the current-voltage (I-V) characteristics of SB-GNRFETs. We carry out accurate approximations of Schottky barrier tunneling, channel charge and current, and which provide improved accuracy while maintaining compactness. This SPICE-compatible compact model surpasses the existing model [15] in accuracy, and enables efficient circuit-level simulations of futuristic GNRFET-based circuits. The proposed model considers various design parameters and process variation effects, including graphene-specific edge roughness, which allows complete and thorough exploration and evaluation of SB-GNRFET circuits. We are able to model both single- and double-gate SB-GNRFETs, so we can evaluate and compare these two types of SB-GNRFET. We also compare circuit-level performance of SB-GNRFETs with multi-gate (MG) Si-CMOS for a scalability study in future generation technology. Our circuit simulations indicate that SBGNRFET has an energy-delay product (EDP) advantage over Si-CMOS; the EDP of the ideal SB-GNRFET (assuming no process variation) is ~1.3% of that of Si-CMOS, while the EDP of the non-ideal case with process variation is 136% of that of Si-CMOS. Finally, we study technology scaling with SBGNRFET and MG Si-CMOS. We show that the EDP of ideal (non-ideal) SB-GNRFET is ~0.88% (54%) EDP of that of Si-CMOS as the technology nodes scale down to 7 nm.

In this paper, many works have been focused on synthesis, verification, and testing of threshold circuits due to the rapid development in efficient implementation of threshold logic circuits. To minimize the hardware cost of threshold circuit implementation, this paper proposes a heuristic that consists of rewiring operations and a simplification procedure. Additionally, a subset of input vectors of a gate, called critical-effect vectors, are proved to be complete for formally verifying the equivalence of two threshold logic gates, instead of the whole truth table in this paper. This achievement can accelerate the equivalence checking of two threshold logic gates. The experimental results show that the proposed heuristic can efficiently reduce the cost.

Power consumption has become one of the primary challenges to meet the Moore’s law. For reducing power consumption, Single-Transistor-Equalizer (SET) at room temperature has been demonstrated as a promising device for extending Moore’s law due to its ultra-low power consumption during operation. Prior work has proposed an automated mapping approach for SET arrays which focuses on minimizing the number of hexagons in a SET array. However, the area of a SET array is more related to the width. Consequently, in this work, we propose an approach for width minimization of the SET arrays. The experimental results show that the proposed approach saves 26% of width compared with the state-of-the-art for a set of MCNC and IWLS 2005 benchmarks while spending similar CPU time.

As fabrication processes exploit even deeper submicron technology, power dissipation has become a crucial issue for most electronic circuit and system designs nowadays. In particular, leakage power is becoming a dominant source of power consumption. Recently, the reconﬁgurable single-electron transistor (SET) array has been proposed as an emerging circuit design style for continuing Moore’s law due to its ultra-low power consumption. Several automated synthesis approaches have been developed for the reconﬁgurable SET array in the past few years. Nevertheless, all of those existing methods consider fabrication constraints, which are mandatory, merely in late synthesis stages. In this paper, we propose a synthesis algorithm, featuring both variable reordering and product term reordering, for area minimization. In addition, our algorithm takes those mandatory fabrication constraints into account in early stages for better outcomes. Experimental results show that our new method can achieve an area reduction of up to 24% as compared to current state-of-the-art techniques.

The realization of large-scale quantum computing is no longer simply a hardware question. The rapid development of quantum technology has resulted in dozens of control and programming problems that should be directed towards the classical computer science and engineering community. One such problem is known as Pauli tracking. Methods for implementing quantum algorithms that are compatible with usual error correction technology utilize extensive quantum teleportation protocols. These protocols are intrinsically probabilistic and result in correction operators that occur as byproducts of teleportation. These byproduct operators do not need to be corrected in the quantum hardware itself, but are tracked through the circuit and output results reinterpreted. This tracking is routinely ignored in quantum information as it is assumed that tracking algorithms will eventually be developed. In this work we help fill this gap and present an algorithm for tracking byproduct operators through a quantum computation.

5.7 Test generation and optimization

Chair: Xiaojing Wen, Kyushu Institute of Technology, JP; Grzegorz Mrugalski, Mentor Graphics, PL

Efficient SMT-based ATPG for Interconnect Open Defects
Dominik Ehr, Karsten Schreiber, Matthias Sauer, Bernd Becker

Interconnect opens are known to be one of the predominant defects in nanoscale technologies. However, automatic test pattern generation for open faults is challenging, because of their rather unstable behaviour and the numerous electric parameters which need to be considered. Thus, most approaches try to avoid accurate modeling of all constraints and use simplified fault models in order to detect as many faults as possible or make assumptions which decrease both complexity and accuracy. This paper presents a new SMT-based approach which for the first time supports the Robust Enhanced Aggressor Victim model without restrictions and handles occlusions. It is combined with the first open fault simulator fully supporting the Robust Enhanced Aggressor Victim model and thereby accurately considering unknown values. Experimental results show the high efficiency of the new method outperforming previous approaches by up to two orders of magnitude.

Index Terms - Interconnect opens, test generation, ATPG, unknown values, SMT

Interconnect Test for 3D Stacked Memory-on-Logic
Mohdshahla Taouil, Mahmoud Masadah, Saif Hamdouli, Erik Jan Marmansen

Three-dimensional stacked IC (3D-SIC) technology based on Through-Silicon Vias (TSVs) provides numerous advantages as compared to traditional 2D-ICs. A potential application is memory stacked on logic, providing enhanced throughput, reduced latency and power consumption. However, testing the TSV interconnects between the two dies is challenging, as both the memory and the logic die might come from different manufacturers. Currently, no standard exists and the proposed solutions fail to address dynamic and time-critical faults (at speed testing). In addition, memory vendors have not been in favor to put additional DfT structures such as JTAG for interconnect testing on their memory devices. This paper proposes a new Memory Based Interconnect Test (MBIT) approach for 3D stacked memories. Our test patterns are applied by read and write instructions to the memory and are validated by a case study where a 3D memory is assumed to be stacked on a MIPS64 processor. The main benefits of the MBIT approach are: (1) zero area overhead, (2) the ability to detect both static and dynamic faults and perform at speed testing, (3) flexibility in applying any test pattern, as this can be executed by the CPU on the logic die and (4) extreme short test execution time.

An Effective Approach to Automatic Functional Processor Test Generation for Small-Delay Faults
Yi-Hang Chen, Jian-Yu Chen, Juinn-Dar Huang

Keywords - single-electron transistor; automatic synthesis; reconﬁgurable; area minimization; binary decision diagram.
Functional microprocessor test methods provide several advantages compared to DFT approaches, like reduced chip cost and at-speed execution. However, the automatic generation of functional test patterns is an open issue. In this work we present an approach for the automatic generation of functional microprocessor test sequences for small-delay faults based on Bounded Model Checking. We utilize an ATPG framework for small-delay faults in sequential, non-scan circuits and propose a method for constraining the input space for generating functional test sequences (i.e., test programs). We verify our approach by evaluating the miniMIPS microprocessor. In our experiments we were able to reach over 97% fault efficiency. To the best of our knowledge, this is the first fully automated approach to functional microprocessor test for small-delay faults.

Keywords: Test; analog/RF integrated circuits; alternate test, specification prediction, test confidence

We have production test data from 10,000 devices.

The rise of dark silicon is driving a new class of architectural techniques that “spend” area to “buy” energy efficiency. In this talk we examine two new frameworks employed by computer architects to understand the challenges and opportunities that await us. The first is the utilization war [9], a simple model that architects use to understand how technology scaling under post-Dennard assumptions will affect hardware design. The second framework is the four horsemens taxonomy [2] that compiles four key approaches that future chip designers will use to attack the dark silicon problem. We describe recent research projects that typify these approaches, including GreenDroid [6], a massively heterogeneous 28 nm processor being developed at UCSB. Finally, we conclude with some directions (and non-directions) that the human brain could offer for refactoring the computational stack for dark silicon [1].

Functional microprocessor test methods provide several advantages compared to DFT approaches, like reduced chip cost and at-speed execution. However, the automatic generation of functional test patterns is an open issue. In this work we present an approach for the automatic generation of functional microprocessor test sequences for small-delay faults based on Bounded Model Checking. We utilize an ATPG framework for small-delay faults in sequential, non-scan circuits and propose a method for constraining the input space for generating functional test sequences (i.e., test programs). We verify our approach by evaluating the miniMIPS microprocessor. In our experiments we were able to reach over 97% fault efficiency. To the best of our knowledge, this is the first fully automated approach to functional microprocessor test for small-delay faults.

Keywords: Test; analog/RF integrated circuits; alternate test, specification prediction, test confidence

We have production test data from 10,000 devices.
6.3 Management of Micro/Macro Renewable Energy Storage Systems

Chair: Geoff Merrett, University of Southampton, UK; Davide Brunelli, University of Trento, Italy

Impact of Steep-Slope Transistors on Non-Von Neumann Architectures: CNN Case Study

Indranil Palit, Behnam Sedighi, Andras Horvath, X. Sharon Hu, Joseph Nahas, Michael Niemier

A Cellular Neural Network (CNN) is a high-performance analog processor that can significantly outperform von Neumann architectures for certain classes of problems. Here, we show how emerging, beyond-CMOS devices could help to further enhance the capabilities of CNNS, particularly for solving problems with non-binary outputs. We show how CNNs based on devices such as graphene transistors - with steep current-growth regions separated by negative differential resistance (NDR) in their I-V characteristics - could be used to realize multiple patterns simultaneously. This would allow multiple CNN architectures to be implemented in a single chip, thereby achieving higher energy efficiency.


Real-time Optimization of the Battery Banks Lifetime in Hybrid Residential Electrical Systems

Maurizio Rossi, Alessandro Toppano, Davide Brunelli

We present a real-time optimization framework to manage Hybrid Residential Electrical Systems (HRES) with multiple Energy sources and heterogeneous storage units. The optimization is performed by solving the nonlinear optimization problem with respect to the voltage, current and power levels of the battery banks. The optimization strategy aims at maximizing the lifetime of the battery banks and at reducing the energy bill by managing the variability of the PV source, in price-varying scenarios. We use a Dynamic Programming (DP) algorithm to schedule the use of the PV-based energy production and the depth-of-discharge (DoD) of the battery banks in order to minimize the number of cycles and the total energy cost. The proposed optimization strategy takes into account the variability of the PV source, in price-varying scenarios. This allows for a more efficient utilization of the renewable energy sources and reduces the cost of electricity.


Optimal Dimensioning of Active Cell Balancing Architectures

Swaminathan Narayanasamy, Sebastian Steinhorst, Martin Lukasiewicz, Matthias Kauer, Samayth Chakraborty

This paper presents an approach to optimal dimensioning of active cell balancing architectures, which are of increasing relevance in Electrical Energy Storages (EESs) for Electric Vehicles (EVs) or stationary applications such as smart grids. Active cell balancing equalizes the state of charge of cells within a battery pack via charge transfers, increasing the effective capacity and reducing the risk of cell aging and failure. The proposed approach integrates MPSoC with an on-chip microfluidic fuel cell network for joint cooling and power supply (i.e., localized power generation and delivery). By providing alternative means to power delivery integrated with cooling, MPSoCs are expected to gain in I/O connectivity. Based on this disruptive technology, we envision the removal of the current limits of power delivery and heat dissipation in MPSoC designs, subsequently avoiding dark silicon and enabling a paradigm shift in future energy-proportional computing architecture designs.
and lifetime. While optimization approaches have been introduced into the design process of several aspects of EESIs, active cell balancing architectures have, until now, been systematically optimized in terms of their components. Therefore, this paper analyzes existing architectures to develop design metrics for energy dissipation, installation volume, and balancing current. Based on these design metrics, a methodology to efficiently obtain Pareto-optimal configurations for a wide range of inductors and transistors at different balancing currents is developed. Our methodology is then applied to a case study, optimizing two state-of-the-art architectures using realistic balancing algorithms. The results give evidence of the applicability of systematic optimization in the domain of cell balancing, leading to higher energy efficiencies with minimized installation space.

[149] Optimal Design and Management of a Smart Residential PV and Energy Storage System
Dx Zhu, Yanzhi Wang, Naehyuck Chang, Massoud Pedram

Solar photovoltaic (PV) technology has been widely deployed in large power plants operated by utility companies. However, the home owners are not yet convinced of the saving cost benefits of this technology, and consequently, in spite of government subsidies, they have been reluctant to install PV systems in their homes. The main reason for this is the absence of a complete and truthful analysis which could explain to home owners what conditions spending money on a PV system can actually save them money over a long-term, but known, time horizon. This paper thus presents a design and management mechanism for a smart residential energy system comprising PV modules, electrical energy storage banks, and connection circuits connected to the power grid. First, we figure out how much savings can be achieved by a system with given PV modules and EES bank capacities by optimally solving the daily energy flow control problem of such a system. Based on the daily optimization results, we come up with the optimal system specifications with a fixed budget. Experiments are conducted for various electricity prices and different profiles of PV output power and load demand. Results show that the designed system breaks even in 6 years and in the system lifetime achieves up to 8% annual profit besides paying back the budget.

Enrico Macelli, Ningning Wang, Sabahl Roy, Michael Hayes, Rudi Paolo Paganeli, Marco Tarantini, Aldo Romani

This paper presents a design study of a new topology for miniaturized bondwire transformers fabricated and assembled with standard IC bonding wires and toroidal ferrite (Fair-Rite 5975000081) as a magnetic core. The microtransformer realized on a PCB substrate, enables the build of magnetics on-top-of-chip, thus leading to the design of high power density components. Impedance measurements in a frequency range between 100 kHz to 5 MHz, show that the secondary self-inductance is enhanced from 0.3 µH with an epoxy core to 315 µH with the ferrite core. Moreover, the micromachined ferrite improves the coupling coefficient from 0.1 to 0.9 and increases the effective turns ratio from 0.5 to 35. Finally, a low-voltage IC DC-DC converter solution, with the transformer mounted on-top, is proposed for energy harvesting applications.

Keywords - bonding wires; bondwire; converter; DC-DC; energy harvesting; ferrite; magnetics; on-chip; PCB; transformer.

[151] Providing Regulation Services and Managing Data Center Peak Power Budgets
Bans Aksanli, Tajana Rosing

Data centers are good candidates for providing regulation services in the power markets due to their large power consumption and flexibility. In this paper, we develop a framework that explores the feasibility of data center participation in these markets. We use a battery-based design that can not only help with providing ancillary services, but can also limit peak power costs without any workload performance degradation. The results of our study using data for a 21MW data center show up to $480,000/year savings can be obtained, corresponding to 1280 more servers providing services.

[152] The Energy Benefit of Level-crossing Sampling Including the Actuator's Energy Consumption
Burkhard Hensel, Klaus Kabbtsch

When using level-crossing (also called send-on-delta) sampling in control loops, messages can be saved compared to periodic sampling without degrading control performance. While it is clear that reducing messages improves also the energy efficiency of battery-powered sensor devices, this can be disadvantageous for the energy efficiency the actuator device. This paper addresses the question, under which conditions level-crossing sampling is also for the actuator device more energy efficient than periodic sampling. It is shown that there is an optimum inter-sample interval. Methods for reaching this optimum by appropriate controller and transmission settings are given. The theory is demonstrated using several known, standardized wireless network protocols.

6.4 Power Delivery and Distribution

Chair: Edith Beigné, CEA LETI Grenoble, FR; Domenik Heims, OFFIS Oldenburg, DE

Masaaki Kondo, Hikoaki Kobayashi, Ryuchii Sakamoto, Motoki Mida, Jun Tsukamoto, Mita Namiki, Weihan Wang, Hideo Amano, Kensaku Matsunaga, Masaru Kudo, Kimiyoshi Usami, Toshiya Komoda, Hiroshi Nakamura

Power-performance efficiency is still remaining a primary concern for microprocessor designers. One of the sources of power inefficiency for recent LSI chips is increasing leakage power consumption. Power-gating is a well known technique to reduce leakage power consumption by switching off the power supply to idle logic blocks. Recently, fine-grained power-gating is emerged as a technique to minimize leakage current during the active processor cycles by switching on and off a logic blocks in much finer temporal/spatial granularity. Though fine-grained power-gating is useful, a comprehensive evaluation and analysis has not been conducted on a real LSI chips. In this paper, we evaluate fine-grained run-time power-gating for microprocessors' functional units using a real embedded microprocessor. We also introduce an architecture and compiler co-operative power-gating scheme which mitigates negative power reduction caused by the energy overhead associated with fine-grained power-gated gates. The experimental results with a fabricated core shows that a hardware-based scheme saves power consumption of functional units by 44% and hardware compiler co-operative scheme further improves power efficiency by 5.9% when core temperature is 25°C.

[154] SuperRange: Wide Operational Range Power Delivery Design for Both STV and NTV Computing
Xin He, Guifai Yan, Yinhe Han, Xiaowei Li

The load power range of modern processors is greatly enlarged because many advanced power management techniques like dynamic voltage frequency scaling, Turbo boosting, and Near Threshold Voltage technologies are now available for microprocessors designers. One of the sources of power inefficiency for recent LSI chips is increasing leakage power consumption. Power-gating is a well known technique to reduce leakage power consumption by switching off the power supply to idle logic blocks. Recently, fine-grained power-gating is emerged as a technique to minimize leakage current during the active processor cycles by switching on and off a logic blocks in much finer temporal/spatial granularity. Though fine-grained power-gating is useful, a comprehensive evaluation and analysis has not been conducted on a real LSI chips. In this paper, we evaluate fine-grained run-time power-gating for microprocessors' functional units using a real embedded microprocessor. We also introduce an architecture and compiler co-operative power-gating scheme which mitigates negative power reduction caused by the energy overhead associated with fine-grained power-gated gates. The experimental results with a fabricated core shows that a hardware-based scheme saves power consumption of functional units by 44% and hardware compiler co-operative scheme further improves power efficiency by 5.9% when core temperature is 25°C.

[155] SuperRange: Wide Operational Range Power Delivery Design for Both STV and NTV Computing
Xin He, Guifai Yan, Yinhe Han, Xiaowei Li

The load power range of modern processors is greatly enlarged because many advanced power management techniques like dynamic voltage frequency scaling, Turbo boosting, and Near Threshold Voltage technologies are now available for microprocessors designers. One of the sources of power inefficiency for recent LSI chips is increasing leakage power consumption. Power-gating is a well known technique to reduce leakage power consumption by switching off the power supply to idle logic blocks. Recently, fine-grained power-gating is emerged as a technique to minimize leakage current during the active processor cycles by switching on and off a logic blocks in much finer temporal/spatial granularity. Though fine-grained power-gating is useful, a comprehensive evaluation and analysis has not been conducted on a real LSI chips. In this paper, we evaluate fine-grained run-time power-gating for microprocessors' functional units using a real embedded microprocessor. We also introduce an architecture and compiler co-operative power-gating scheme which mitigates negative power reduction caused by the energy overhead associated with fine-grained power-gated gates. The experimental results with a fabricated core shows that a hardware-based scheme saves power consumption of functional units by 44% and hardware compiler co-operative scheme further improves power efficiency by 5.9% when core temperature is 25°C.

6.5 Beyond EDA: Extending the Application Domain of Formal Methods

Chair: Christoph Scholl, University of Freiburg, DE; Gianpiero Cabodi, Politecnico di Torino, IT

Samantak Gangopadhyay, Youngtak Lee, Saad Bin Nasser, Anil Raychowdhury

Discreet time digital linear regulators, including low dropout regulators (LDOs) have become competitive in multi-VCC digital systems for fine-grained spatio-temporal voltage regulation and distribution. However, wide dynamic current range of the digital load circuits poses serious problems in maintaining stability and high efficiency at all corners. In this paper we present a control model for discrete time LDOs and demonstrate how online adaptive control can be employed for consistent performance and high efficiency across the load current range.
In this paper we present MaxBMC, a novel formalism for solving optimization problems in sequential systems. Our approach combines techniques from symbolic SAT-based Bounded Model Checking (BMC) and incremental MaxSAT, leading to the first MaxBMC solver. In traditional BMC safety and liveness properties are validated. We extend this formalism: in case the required property is satisfied, an optimization problem is defined to maximize the quality of the reached witnesses. Further, we compare its qualities in different depths of the system, leading to Pareto-optimal solutions. We state a sound and complete algorithm that not only tackles the optimization problem but moreover verifies whether a global optimum has been identified by using a complete BMC solver as back-end. As a first reference application we present the problem of circuit initialization. Additionally, we give pointers to other tasks which can be covered by our formalism quite naturally and further demonstrate the efficiency and effectiveness of our approach.

For effectively solving quantified boolean formulas (QBFs), preprocessors have shown to be of great value. A preprocessor rewrites a formula such that helpful information is made explicit and irrelevant information is removed. For this purpose, techniques, which would be too costly when repeatedly applied during the solving process, are used. Unfortunately, most preprocessing techniques are not model preserving and therefore incompatible with certification frameworks. In consequence, the application of a preprocessor prohibits the extraction of witnesses encoding a solution or a counterexample of a formula. In this paper, we show how to obtain partial witnesses from preprocessed QBFs. Partial witnesses are assignments for the variables of the outermost quantifier block and are extensible to full witnesses, which are usually represented as functions reflecting the dependencies between variables. For many applications, however, partial witnesses are sufficient. We modified the publicly available preprocessor bloqqer for extracting partial witnesses. We empirically compare the effectiveness of the modified and the original version of bloqqer. Further, we apply the new version of bloqqer for solving hardware synthesis problems for which it turns out to be extremely beneficial.

Function pipelining is a key transformation in behavioral synthesis. However, synthesizing the complex pipeline logic is an error-prone process. Sequential equivalence checking (SEC) support is highly desired to provide confidence in the correctness of synthesized pipelines. However, SEC for function pipelining is challenging due to the significant difference between the behavioral specification and synthesized RTL. Furthermore, function pipelines include hardware logic for dynamically inserting “bubbles” (pipeline stalls), which bring additional difficulties in equivalence checking. We develop an SEC framework for behaviorally synthesized function pipelines by (1) building a reference pipeline model with a certified function pipelining transformation, which faithfully captures bubbles insertion; and (2) checking the equivalence between the reference model and synthesized RTL. We demonstrate the scalability of our approach on industry-strength designs synthesized by a commercial tool.

Recent developments in the fabrication technology attracted the attention of optical engineers and physicists in the area of VLSI photonics. Due to the physical nature of light-wave systems and their usage in safety critical domains such as human surgeries and high budget space missions, it is indispensable to build high assurance systems. Traditionally, the analysis of such systems has been carried out by paper-and-pencil based proofs and numerical computations. However, these techniques cannot provide perfectly accurate results due to the risk of human error and inherent approximations of numerical algorithms. In order to overcome these limitations, we propose to use higher-order logic theorem proving to improve the analysis in the domain of integrated optics or VLSI photonics. In particular, this paper provides a higher-order logic formalization of optical microresonators which are the most fundamental building blocks of many photonic devices. In order to illustrate the practical utilization of our work, we present the formal analysis of 2-D microresonator lattice optical filters.

Despite the progress of high-level languages and tools, Register Transfer Level (RTL) is still far by the dominant input format for high performance digital designs. Experienced designers can directly express their microarchitectural intuitions in RTL. Yet, RTL is terribly verbose, burdened with trivial details, and thus error prone. In this paper, we augment a modern RTL language (Chisel) with new semantic elements to express an imprecise specification: a sketch. We show how, in combination with a naïve, unoptimized, but functionally correct reference, a designer can directly express their microarchitectural intuitions in RTL. Yet, RTL is terribly verbose, burdened with trivial details, and thus error prone. In this paper, we augment a modern RTL language (Chisel) with new semantic elements to express an imprecise specification: a sketch. We show how, in combination with a naïve, unoptimized, but functionally correct reference, a designer can directly express their microarchitectural intuitions in RTL. Yet, RTL is terribly verbose, burdened with trivial details, and thus error prone. In this paper, we augment a modern RTL language (Chisel) with new semantic elements to express an imprecise specification: a sketch. We show how, in combination with a naïve, unoptimized, but functionally correct reference, a designer can directly express their microarchitectural intuitions in RTL.
original synchronous model, i.e., for each variable exactly the same flow of data values must be observed and only the membership to synchronous reaction steps is no longer explicitly given. In this paper, we prove that this correctness guarantee is given provided that (1) each component knows which of the input values have to be used for the next reaction (endochrony), (2) each component is able to perform the reaction (constructiveness), and (3) components agree on the clocks of their shared variables (isochrony/clock-consistency).

[164] Tightly-Coupled Hardware Support to Dynamic Parallelism Acceleration in Embedded Shared Memory Clusters
Pacito Burgio, Giuseppe Taglavini, Francesco Conti, Andrea Marongiu, Luca Benvi

Modern designs for embedded systems are increasingly embracing cluster-based architectures, where small sets of cores communicate through tightly-coupled shared memory banks and high-performance interconnections. At the same time, the complexity of modern applications requires new programming abstractions to exploit dynamic and/or irregular parallelism on such platforms. Supporting dynamic parallelism in systems which i) are resource-constrained and ii) run applications with small units of work calls for a runtime environment which has minimal overhead for the scheduling of parallel tasks. In this work, we study the major sources of overhead in the implementation of OpenMP dynamic loops, sections and tasks, and propose a hardware implementation of a generic Scheduling Engine (WHSE) which fits the semantics of the three constructs. The WHSE is designed as a tightly-coupled block to the PEiS within a multi-core cluster, communicating through a shared-memory interface. This allows very fast programming and synchronization with the controlling PEiS, fundamental to achieving fast dynamic scheduling, and ultimately to enable fine-grained parallelism. We prove the effectiveness of our solutions with real applications and synthetic benchmarks, using a cycle-accurate virtual platform.

[165] p-OFTL: An Object-based Semantic-aware Parallel Flash Translation Layer
Wei Wang, Youyou Lu, Jiewu Shu

With increased density and decreased price, flash memory has been widely used in storage systems for its low latency and low power features. However, traditional storage systems are designed and excessively optimized for magnetic disks, and the potential of flash memory is not brought into full play in the form of Solid State Drives (SSDs). In this paper, we propose p-OFTL, an object-based semantic-aware parallel flash translation layer (FTL). p-OFTL removes the mapping table in the FTL and directly manages the flash memory in file objects, which enables optimization of data layout in the flash using object semantics. While the removing of the mapping table improves system performance, a challenge remains to exploit the internal parallelism when maintaining the continuity of logical addresses in each object, which is essential for efficient garbage collection. To address this challenge, p-OFTL statically remaps the addresses by shifting the bits in the addresses, which spreads writes to different internal parallel units without another mapping table. Also, p-OFTL employs a semantic-aware data grouping algorithm to group data pages by trading off the hot-cold clustering for the continuity of logical addresses, so as to reduce the page movement in garbage collection. Experiments show that p-OFTL improves system performance by 4.0% - 10.3% and reduces garbage collection overhead by 15.1% - 32.5% in semantic-aware data grouping compared to those in semantic-unaware data grouping algorithms.

[166] Using Guided Local Search for Adaptive Resource Reservation in Large-scale Embedded Systems
Timon D. ter Braak

To maintain a predictable execution environment, an embedded system must ensure that applications are, in advance, provided with sufficient resources to process tasks, exchange information and to control peripherals. The problem of assigning tasks to processing elements with limited resources, and routing communication channels through a capacitated interconnect is combined into an integer linear programming formulation. We describe a guided local search algorithm to solve this problem at run-time. This algorithm allows for a hybrid strategy where configurations computed at design-time may be used as references to lower the computational overhead at runtime. Computational experiments on a dataset with 100 tasks and 20 processing elements show the effectiveness of this algorithm compared to state-of-the-art solvers CPLEX and Gurobi. The guided local search algorithm finds an initial solution within 100 milliseconds, is competitive for small platforms, scales better with the size of the platform, and has lower memory usage (2-19%).

[167] Accelerating Graph Computation with Racetrack Memory and Pointer-Assisted Graph Representation
Euyihck Park, Sungjoo Yoo, Sunggu Lee, Helen Li

The poor performance of NAND Flash memory, such as long access latency and large granularity access, is the major bottleneck in graph processing. This paper proposes an intelligent storage for graph processing which is based on fast and low cost racetrack memory and a pointer-assisted graph representation. Our experiments show that the proposed intelligent storage based on racetrack memory reduces total processing time of three representative graph computations by 40.2% - 86.9% compared to the graph processing, GraphChi, which exploits sequential accesses based on normal NAND Flash memory-based SSD. Faster execution also reduces energy consumption by 39.6% - 90.0%. The in-storage processing capability gives additional 10.5% - 16.4% performance improvements and 12.0% - 14.4% reduction of energy consumption.

6.7 Hardening Approaches at Different Design Levels

[168] Nostradamus: Low-Cost Hardware-Only Error Detection for Processor Cores
Ralph Nathan, Daniel J. Sorin

We propose a new, low-cost, hardware-only scheme to detect errors in superscalar, out-of-order processor cores. For each instruction decoded, Nostradamus compares what the instruction is expected to do against what the instruction actually does. We implement Nostradamus in RTL on top of a baseline superscalar, out-of-order core, and we experimentally evaluate its ability to detect injected errors. We also evaluate Nostradamus's area and power overheads.

[169] Word-Line Power Supply Selector for Stability Improvement of Embedded SRAMs in High Reliability Applications
B. Ablaza, C. Camorna, S. Bota

Embedded SRAM yield dominates the overall ASIC yield, therefore the methodologies centered on improving SRAM cell stability will be introduced in the design as a mandatory. Word-line voltage modulation has shown that it is possible to improve cell stability during access operations. The high variability of physical and performance parameters introduce the need to adopt adaptive solutions to adequately improve SRAM cell stability. In this work, we present a wordline voltage selector circuit designed to modulate power-supply word-line voltage at each individual embedded SRAM block. The final area overhead is minimal and several strategies can be implemented with the embedded SRAM allowing adjust wordline voltage during the life of ASIC, taking into account different operation, aging and degradation effects.

Keywords: SRAM stability, Word-line modulation, High Reliability applications.

[170] A High Performance SEU-Tolerant Latch for Nanoscale CMOS Technology
Zhengfeng Huang

This paper presents a high performance latch to tolerate radiation-induced single event upset in 45 nm CMOS technology. The latch can improve robustness by masking the soft errors utilizing Muller C-element and dual modal redundancy hardening. The power dissipation, propagation delay and reliability of the presented SEU-tolerant latch are analyzed by SPICE simulations. The results show that the presented latch provides a higher robustness and lower power-delay product than classical implementations and alternative hardened solutions.

Index Terms: Transient fault, soft error, single event upset, hardened-by-design, C-element.

[171] A Low-Cost Radiation hardened Flip-flop
Yang Lin, Mark Zwolinski, Basel Nahak

The aggressive scaling of semiconductor devices has caused a significant increase in the soft error rate caused by radiation hits. This has led to an increasing need for fault-tolerant techniques to maintain system reliability. Conventional radiation hardening techniques, typically used in safety-critical applications, are prohibitively expensive for non-safety-critical
electronics. This work proposes a novel flip-flop architecture named SETOFF which significantly improves circuit resilience to radiation hits over previous techniques. In addition, compared to other techniques such as a TMR latch, SETOFF reduces the area and performance overheads by up to 50% and 80%, respectively; the power consumption overhead is also reduced by up to 85%. In addition, a novel reliability metric called radiation-induced failure rate is developed which can be a valuable tool to predict the impact of radiation hits and quantitatively compare the reliability of various radiation hardened techniques. Our analysis shows that the proposed technique can achieve zero SEE failure rate, and significantly reduce the SET failure rate.

Keywords - Soft error, reliability, single-event upset, single-event transient, fault-tolerance.

[172] PSP-Cache: A Low-Cost Fault-Tolerant Cache Memory Architecture
Hamed Farbeh, Seyed Ghassam Mermadi
Cache memories constitute a large fraction of processor chip area and are highly vulnerable to soft errors caused by energetic particles. To protect these memories, most of the modern processors employ Error Detection Codes (EDCs) or Error Correction Codes (ECCs). ECCs impose significant overheads in terms of area and energy; these overheads increase as a function of interleaving EDCs/ECCs to detect/correct multiple errors. This paper proposes a new cache architecture to minimize the area and energy overheads of EDCs/ECCs in setassociative L1 caches. Simulation results for a 4-way setassociative cache show that the proposed architecture reduces both the area and static power overheads of parity code by about 75% and the dynamic energy overhead by about 73% in comparison to conventional cache architecture. These reduction figures are about 68% and about 66%, respectively, for SECDED code. The above reductions are achieved without affecting the error coverage.

Keywords - Cache Memory, Error Detection and Correction; Soft Errors; Fault Tolerance

[173] A Hybrid Non-Volatile SRAM Cell with Concurrent SEE Detection and Correction
Piln Junsangsri, Fabrizio Lombardi, Jie Han
This paper presents a hybrid non-volatile (NV) SRAM cell with a new scheme for SEE tolerance. The proposed NVSRAM cell consists of a 6T SRAM core and a Resistive RAM (RRAM), made of a 1T and a Programmable Metallization Cell (PMC). The proposed cell has concurrent error detection (CED) and correction capabilities; CED is accomplished using a dual-rail checker, while correction is accomplished by utilizing the restore operation; data from the non-volatile memory element is copied back to the SRAM core. The dual-rail checker utilizes two XOR gates each made of 2 inverters and 2 ambipolar transistors, hence, it has a hybrid nature. Extensive simulation results are provided. The simulation results show that the proposed scheme is very efficient in terms of numerous figures of merit such as delay and circuit complexity and thus applicable to integrated circuits such as FPGAs requiring secure on-chip non-volatile storage (i.e. LUT6) for multi-context configurability.

Keywords - Memory Cell, Programmable Metallization

7.0 Special Day Keynote

[174] The Connected Car and Its Implication to the Automotive Chip Roadmap
Michael Bolle
The automotive industry is in a radical change process driven by technology. On the one hand side the proliferation of communication technologies into the car leads to internet connected vehicles. The vehicles then become an integral part of the internet - opening new processing paradigms for the car itself. On the other hand the vehicle itself significantly expands its sensor and processing capabilities by the use of radar, video, ultrasound sensors and usage of state of the art GPU and CPU processor architectures. In our talk we will address both developments and outlook foresee future applications as future driving assistant and infotainment systems as well as highly automated driving. We will discuss major requirements for the future electrical architectures and implications for future automotive chips.

7.2 Embedded Tutorial: Cross Layer Resiliency in Real World

Organizer: Vikas Chandra, ARM, US
Chairs: Yanjing Li, Intel, US; Ulf Schlichtmann, TUM, DE

[175] Cross Layer Resiliency in Real World
Vikas Chandra
Resilience at different design hierarchies will be needed in Complex SoCs to handle failures due to variability, reliability and design errors (logical or electrical). The main reasons for the marginal behavior are shear design complexity, manufacturing processes, temporal variability and operating conditions. In this session, we will cover the basics of cross layer resiliency and explore the reliability challenges in both embedded processors as well as large scale computing resources.

7.3 Low Power Methods and Multicore Architectures for Mobile Health Applications

Chairs: Ashoka Sathanur, Philips, NL; Giovanni Ansaldi, EPFL, CH

[176] Hardware/Software Approach for Code Synchronization in Low-Power Multi-Core Sensor Nodes
Rubén Brasjos, Ahmed Dogan, Ivan Benetta, Giovanni Ansaldi, David Atienza
Latest embedded bio-signal analysis applications, targeting low-power Wireless Body Sensor Nodes (WBSNs), present conflicting requirements. On one hand, bio-signal analysis applications are continuously increasing their demand for high computing capabilities. On the other hand, long-term signal processing in WBSNs must be provided within their highly constrained energy budget. In this context, parallel processing effectively increases the power efficiency of WBSNs, but only if the execution can be properly synchronized among computing elements. To address this challenge, in this work we propose a hardware/software approach to synchronize the execution of bio-signal processing applications in multi-core WBSNs. This new approach requires little hardware resources and very few adaptations in the source code. Moreover, it provides the necessary flexibility to execute applications with an arbitrarily large degree of complexity and parallelism, enabling considerable reductions in power consumption for all multi-core WBSN execution conditions. Experimental results show that a multi-core WBSN architecture using the illustrated approach can obtain energy savings of up to 40%, with respect to an equivalent singlecore architecture, when performing advanced bio-signal analysis.


[177] Hybrid Memory Architecture for Voltage Scaling in Ultra-Low Power Multi-Core Biomedical Processors
Daniele Bortolotti, Andrea Bartolini, Christian Weiss, Davide Rossi, Luca Benini
Technology scaling enables today the design of sensor-based ultra-low cost chips well suited for emerging applications such as wireless body sensor networks, urban life and environment monitoring. Energy consumption is the key limiting factor of this up-coming revolution and memories are often the energy bottleneck mainly due to leakage power. This paper proposes an ultra-low power multi-core architecture targeting eHealth monitoring systems, where applications involve collection of sequences of slow biomedical signals and highly parallel computations at very low voltage. We propose a hybrid memory architecture that combines 6T SRAM and 8T SRAM operating in the same voltage domain and capable of dispatching at high voltage a normal operation and at low voltage a fully reliable small memory partition (8T) while the rest of the memory (8T) is state-retentive. Our architecture offers significant energy savings with a low area overhead in typical eHealth Compressed Sensing based applications.

[178] Context Aware Power Management for Motion-sensing Body Area Network Nodes
Fulvio Casamasima, Eliazaeta Farita, Luca Benini
Body Area Networks (BANs) are widely used mainly for healthcare and fitness purposes. In both cases, the lifetime of sensor nodes included in the BAN is a key aspect that may affect the
functionality of the whole system. Typical approaches to power management are based on a trade-off between the data rate and the monitoring time. Our work introduces a power management layer capable to opportunistically use data sampled by sensors to detect contextual information such as user activity and adapt the node operating point accordingly. The use of this layer has been demonstrated on a commercial sensor node, increasing its battery lifetime up to a factor of 5.

Keywords - body area networks, power management, context-awareness, healthcare applications

[179] A Scalable and Energy-Efficient Approach for Spectral Analysis of Heart Rate Variability
Georgios Karakostantinis, Avinash Sankaranarayanan, Mohamed M. Sabry, David Atienza, Andreas Burg

Today there is a growing interest in the integration of health monitoring applications in portable devices necessitating the development of methods that improve the energy efficiency of such systems. In this paper, we present a systematic approach that enables energy-quality trade-offs in spectral analysis systems for bio-signals, which are useful in monitoring various health conditions as those associated with the heart-rate. To enable such trade-offs, the processed signals are expressed initially in a basis in which significant components that carry most of the relevant information can be easily distinguished from the parts that influence the output to a lesser extent. Such a decomposition facilitates the pruning of operations associated with the less significant signal components leading to power savings with minor quality loss since only less useful parts are pruned under the given requirements. To exploit the attributes of the modified spectral analysis system, thresholding rules are determined and adopted at design- and run-time, allowing the static or dynamic pruning of less-useful operations based on the accuracy and energy requirements. The proposed algorithm is implemented on a typical sensor node simulator and results show up to 82\% energy savings when static pruning is combined with voltage and frequency scaling, compared to the conventional algorithm in which such trade-offs were not available. In addition, experiments with numerous cardiac samples of various patients show that such energy savings come with a 4.9\% average accuracy loss, which does not affect the system-detection capability of sinus arrhythmia which was used as a test case.

[180] Battery Aware Stochastic QoS Boosting in Mobile Computing Devices
Hao Shen, Qiuxen Chen, Qifen Qu

Mobile computing has been woven into everyday lives to a great extent. Their usage is clearly imprinted with user's personal signature. The ability to learn such signature enables immense potential in workload prediction and resource management. In this work, we investigate the user behavior modeling and apply the model for energy management. Our goal is to maximize the quality of service (QoS) provided by the mobile device (i.e., smartphone), while keep the risk of battery depletion below a given threshold. A Markov Decision Process (MDP) is constructed from history user behavior. The optimal management policy is solved using linear programing. Simulations based on real user traces validate that, compared to existing battery energy management techniques, the stochastic control performs better in boosting the mobile devices’ QoS without significantly increasing the chance of battery depletion.

Key words: mobile, battery, energy, Markov Decision Process

7.4 Runtime Memory Optimization and GPU/Manycore Architectures
Chair: Alberto Nannarelli, DTU Copenhagen, DK; Alberto Maci, PolTo Toino, IT

[181] Exploiting Expendable Process-Margins in DRAMs for Run-Time Performance Optimization
Karthik Chandrasekar, Sven Goossens, Christian Wei, Martin Koedam, Benny-Akesson, Norbert Wehrt, Kees Goossens

Manufacturing-time process (P) variations and runtime voltage (V) and temperature (T) variations can affect a DRAMs performance severely. To counter these effects, DRAM vendors provide substantial design-time PVT timing margins to guarantee correct DRAM functionality under worst-case operating conditions. Unfortunately, with technology scaling these timing margins have become large and very pessimistic for a majority of the manufactured DRAMs. While run-time variations are specific to operating conditions and as a result, their margins difficult to optimize, process variations are manufacturing-time effects and excessive process-margins can be reduced at run-time, on a per device basis, if properly identified. In this paper, we propose a generic post-manufacturing performance characterization methodology for DRAMs that identifies this excess in process-margins for any given DRAM device at runtime, while retaining the requisite margins for voltage (V) and temperature (T) variations. By doing so, the methodology ascertains the actual impact of process-variations on the particular DRAM device and optimizes its access latencies (timings), thereby improving its overall performance. We evaluate this methodology on 48 DDR2 devices (from 12 DIMMs) and verify the derived timings under worst-case operating conditions, showing up to 33.3\% and 25.9\% reduction in DRAM read and write latencies, respectively.

[182] Cache Aging Reduction with Improved Performance Using Dynamically Re-sizable Cache
Haron Mahmood, Massimo Poncino, Enrico Maci

Aging of transistors is a limiting factor for long term reliability of devices in sub-100nm technologies. It's a worst-case metric where the lifetime of a device is determined by the earliest failing component. Impact is more serious on memory arrays, where failure of a single SRAM cell would cause the failure of the whole system. Previous works have shown that partitioning based strategies on power management techniques can effectively control aging effects and can extend lifetime of the cache significantly. However, such a benefit comes as a tradeoff with performance which reduces proportionally as the time elapses. To address this problem and provide a single solution to concurrently improve aging, energy and performance of the cache, we propose an architectural solution based on the dynamically re-sizable cache [5] and cache partitioning approaches. By this strategy, the cache is dynamically re-sized and reconfigured whenever a cache block becomes unreliable. Coupling such aging mitigation technique along with dynamically re-sizable cache approach provides on average 30\% lifetime improvement with less than 0.4x degradation in performance whereas, in previous solutions, performance degradation sometimes goes upto 10x.

[183] On GPU Bus Power Reduction with 3D IC Technologies
Young-Joon Lee and Sung Kyu Lim

The complex buses consume significant power in graphics processing units (GPUs). In this paper, we demonstrate how the power consumption of buses in GPUs can be reduced with 3D IC technologies. Based on layout simulations, we found that partitioning and floorplanning of 3D ICs affect the power benefit amount, as well as the technology setup, target clock frequency, and circuit switching activity. For 3D IC technologies using two dies, we achieved the total power reductions of up to 21.5\% over a baseline 2D design.

[184] Process Variation-Aware Workload Partitioning Algorithms for GPUs Supporting Spatial-Multitasking
Paula Aguilara, Jungseob Lee, Amr Farnahi-Farahani, Katherine Morrow, Michael Schulte, Nam Sung Kim

High level programming languages have transformed graphics processing units (GPUs) from domain-restricted devices into powerful compute platforms. Yet many "general-purpose GPU" (GPGPU) applications fail to fully utilize the GPU resources. Executing multiple applications simultaneously on different regions of the GPU (spatial multitasking) thus improves system performance. However, within die process variations lead to significantly different maximum operating frequencies (Fmax) of the streaming multiprocessors (SMs) within a GPU. As the chip size and number of SMs per chip increase, the frequency variation is also expected to increase, exacerbating the problem. The increased number of SMs also provides a unique opportunity: we can allocate resources to concurrently-executing applications based on how those applications are affected by the different available Fmax values. In this paper, we study the effects of per-SM clocking on spatial multitasking-capable GPUs. We demonstrate two factors that affect the performance of simultaneously-running applications: (i) the SM partitioning algorithm that decides how many resources to assign to each application, and (ii) the assignment of SMs to applications based on the operating frequencies of those SMs and the applications characteristics. Our experimental results show that spatial multitasking that partitions SMs based on application characteristics, when combined with per-SM clocking, can greatly improve application performance by up to 46\% on average compared to cooperative multitasking with global clocking.

[185] A Thermally Resilient Integration of Many-core Microprocessors and Main Memory by 2.5D TSI I/Os
Shih-Sian Wu, Kanwen Wang, Sai Manoj P. D., Tsung-Yi Ho, Mingxin Yu, Hao Yu

One memory-logic-integration design platform is developed in this paper with thermal reliability analysis provided for 2.5D through-silicon-interposer (TSI) and 3D through-silicon-via (TSV) based integrations. Temperature-dependent delay and power models have been developed at microarchitecture level for 2.5D and 3D integrations of many-core microprocessors and main memory, respectively. Experiments are performed by general-purpose benchmarks from SPEC CPU2006 and also cloud-oriented benchmarks from Phoenix with the following observations. The memory-logic integration of 3D RC-interconnected TSI I/Os can result in thermal runaway failures due to strong electrical-thermal couplings. On the other hand, the one by 2.5D transmission-line-interconnected TSI I/Os has shown almost the same energy efficiency and better thermal resilience.
Coherent data mining is introduced as a promising architectural technique for reducing cache-to-cache accesses in directory protocol. However, limited on-chip resources cause the accuracy of current prediction to be generally low. Low accuracy would result in a large number of unnecessary or incorrect predictions, which would consequently generate excessive network traffic. This leads to large power and performance overhead for coherent memory access. This paper proposes an early abort mechanism (EBT) that leverages NoC design to reduce the negative effect of wrong prediction operations, thus facilitating overall performance improvement and traffic reduction. Using detailed full-system simulations, we conclude that EBT provides a cost-effective solution for designing efficient multicore processors. To the best of our knowledge, this is the first to leverage on-chip network for the prediction optimization on multicore coherence.

7.5 Emerging Memory Technologies

Chair: Aida Todri, CNRS-JF, Lara Bauer, KIT, DE

[186] Asynchronous Asymmetrical Write Termination (AAWT) for a Low Power STT-MRAM
Rajendra Bishnoi, Mojtaba Ebrahimi, Fabian Oboril, Mehdi B. Tahoori
Spin Transfer Torque (STT) memory is an emerging and promising non-volatile storage technology. However, the high write current is still a major challenge which leads to a huge power consumption of the memory. Due to an inherent torque asymmetry of the Magnetic Tunnel Junction (MTJ) device employed in STT memories, the switching time between parallel and anti-parallel is significantly different. Hence, the write latencies for writing ‘0’ and ‘1’ are also considerably different. In this paper, we propose a technique called Asynchronous Asymmetrical Write Termination (AAWT) which utilizes this asymmetrical behavior to terminate the write operations asynchronously and as a result significantly reduces the write power consumption. Furthermore, we present two different AAWT implementations to determine the actual write termination times. The first one makes use of a clock signal and the second one employs a self-timing approach based on an internal delay element. As shown by our experimental results, AAWT can reduce the total write energy by 30% in average with a negligible area overhead.

[187] Write-Once-Memory-Code Phase Change Memory
Jaylin Li, Kartik Mohanram
This paper describes a write-once-memory-code phase change memory (WOM-code PCM) architecture for next-generation non-volatile memory applications. Specifically, we address the long latency of the write operation in PCM - attributed to PCM SET - by proposing a novel PCM memory architecture that integrates WOM-codes at the memory organization and memory controller levels. The proposed $2\textsuperscript{2/3}$ WOM-code PCM architecture is able to reduce write memory (read) latency by 20.1% (10.2%) on average across general-purpose (SPEC CPU2006), embedded (MBench), and high-performance (SPLASH2) benchmarks. To further improve the write latency of WOM-code PCM, we propose a PCMrefresh approach that uses idle cycles to preemptively set memory PCells to the initial WOM-code state. Results show that WOM-code PCM with PCMrefresh can reduce memory write (read) latency by 54.9% (47.9%) on average across the benchmarks. Finally, to balance write latency improvements against WOM-code PCM overhead, we propose a WOM-code cached PCM (WCPCM) architecture that uses WOM-code PCM as the cache alongside conventional PCM main memory. For just 4.7% memory overhead, WCPCM reduces memory write (read) latency by 47.2% (44.0%) on average across the benchmarks.

[188] Improving STT-MRAM Deniality through Multibit Error Correction
Brandon Del Bel, Jongyeon Kim, Chris H. Kim, Sachin S. Sapatnekar
STT-MRAMs are prone to data corruption due to inadvertent bit flips. Traditional methods enhance robustness at the cost of area/energy by using larger cell sizes to improve the thermal stability of the MTJ cells. This paper employs multibit error correction with DRAM-style refreshing to mitigate errors and provides a methodology for determining the optimal level of correction. A detailed analysis demonstrates that the reduction in nonvolatility requirements afforded by strong error correction translates to significantly lower area for the memory array compared to simpler ECC schemes, even when accounting for the increased overhead of error correction.

[189] Energy Efficient In-Memory AES Encryption Based on Nonvolatile Domain-wall Nanowire
Yuhao Wang, Hao Yu, Dennis Sylvester, Pingtan Kong
The widespread adoption of the AES encryption standard in critical applications has imposed high throughput and low power requirements. Various efforts have been made to address these requirements by using AES encryption on the newly introduced domain-wall nanowire. We show that AES operations can be fully mapped to a low-power, low-latency, and high-performance architecture. The experimental results show that the AES can achieve the best energy efficiency of 24 J/Gb, which is 5x and 6.5x better than CMOS ASIC and memristive CMOL implementations, respectively. Under the same area budget, the proposed DWAES exhibits 6.4x higher throughput and 28% power saving compared to the CMOS ASIC implementation, 1.7x higher throughput and 74% power reduction compared to a memristive CMOL implementation.

Boxion Li, Yu Wang, Yiran Chen, Hai (Helen) Li, Haohong Yang
The emerging neuromorphic computation provides a revolutionary solution to the alternative computing architecture and effectively extends Moore's Law. The discovery of the memristor presents a promising hardware realization of neuromorphic systems with incredible power efficiency, allowing efficiently executing the analog matrix-vector multiplication on the memristor crossbar architecture. However, during computations, the memristor will slowly drift from its initial programmed state, leading to a gradual decline of the computation precision of memristor crossbar-based computing engines (MCE). In this paper, we propose an inline calibration mechanism to guarantee the computation quality of the MCE. The inline calibration mechanism collects the MCE's computation error through 'interrupt-and-benchmark (I&B)' operations and predicts the best calibration time through polynomial fitting of the computation error data. We also develop an adaptive technique to adjust the time interval between two neighbor I&B operations and minimize the negative impact of the I&B operation on system performance. The experimental results demonstrate that the proposed inline calibration mechanism achieves a calibration efficiency of 91.18% on average and negligible performance overhead (δe, 0.439%).
7.6 Performance and Timing Analysis

Chairs: Wang Wang Yi, Uppsala University, SE Petru Petru Eles, Linköping University, SE

[193] Scalability Bottlenecks Discovery in MPSoC Platforms Using Data Mining on Simulation Traces
Sofiane Lagana, Alexandre Témrier, Frédéric Pelrot

Nowadays, a challenge faced by many developers is the profiling of parallel applications so that they can scale over more and more cores. This is especially critical for embedded systems powered by Multi-Processor System-on-Chip (MPSoC), where over-demanding applications have to run smoothly on numerous cores, each with modest power budget. The reasons for the lack of scalability of parallel applications are numerous, and it can be time consuming for a developer to pinpoint the correct one. In this paper, we propose a fully automatic method which detects the instructions of the code which lead to a lack of scalability. The method is based on data mining techniques exploiting low level execution traces produced by MPSoC simulators.

Our experiments show the accuracy of the proposed technique on five different kinds of applications, and how the information reported can be exploited by application developers.

[195] Resource Optimization for CSDF-modeled Streaming Applications with Latency Constraints
Di Liu, Jelena Spasic, Jialu Teddy Zhai, Todor Stefanov, Gang Chen

In this paper, we study the problem of minimizing the number of processors required for scheduling latency-constrained streaming applications modeled as CSDF graphs, where the actors of a CSDF are executed as strictly periodic tasks. We formalize the problem and prove that due to the strict periodicity of actors the problem is an integer convex programming problem, that can be solved efficiently by using an existing convex programming solver. We evaluate our solution approach on a set of 13 real-life streaming applications modeled as CSDF graphs and demonstrate that it can reduce the number of processors in more than 52% of the conducted experiments in comparison to an existing approach.

Baek-Gyu Kim, Hyeon I Hwang, Taepyon Pak, Sang H. Son, Insup Lee

The model-based implementation is to derive an implementation from a model that has been shown to meet requirements. Even though this approach can be used to guarantee that an implementation satisfies functional requirements that are shown to be correct at the model level, it is still challenging to assure timing requirements at the implementation level. We propose a layered approach in testing timing requirements conformance of implemented systems developed by model-based implementation. In our approach, the abstraction boundary of the implemented system is formally defined using Parnas’ four-variables model. Then, the proposed approach tests timing aspects of the interaction between the auto-generated code and the target platform-dependent code based on the four-variables. This approach aims at not only detecting the timing requirement violation, but also at measuring delay-segments that contribute to the timing deviation of the implemented system w.r.t. the model. We show the case study of testing timing requirements of an infusion pump system to illustrate the applicability of the proposed framework.

Kenneth Balck, Olga Grinchtein, Justin Pearson

Within telecommunications development it is vital to have frameworks and systems to replay complicated scenarios on equipment under test, often there are not enough available scenarios. In this paper we study the problem of testing a test harness, which replays scenarios and analyses protocol logs for the Public Warning System service, which is a part of the Long Term Evolution (LTE) 4G standard. Protocol logs are sequences of messages with timestamps; and are generated by different mobile network entities. In our case study we focus on user equipment protocol logs. In order to test the test harness we require that logs have both correct and correct behaviour. It is easy to collect logs from real system runs, but these logs do not show much variation in the behaviour of system under test. We present an approach where we use constraint logic programming (CLP) for both modeling and test generation, where each test case is a protocol log. In this case study, we uncovered previously unknown faults in the test harness.

[199] Time-Decoupled Parallel SystemC Simulation
Jan Henrik Weinstock, Christoph Schumacher, Rainer Leupers, Gerd Ascheid, Laura Tsonaroti

With increasing system size and complexity, designers of embedded systems face the challenge of efficiently simulating these systems in order to enable target specific software development and design space exploration as early as possible. Today’s multicore workstations offer enormous computational power, but traditional simulation engines like the OSCI SystemC kernel only operate on a single thread, thereby leaving a lot of computational potential unused. Most modern embedded system designs include multiple processors. This work proposes SCope, a SystemC kernel that aims at exploiting the inherent parallelism of such systems by simulating the processors on different threads. A lookahead mechanism is employed to reduce the required synchronization between the simulation threads, thereby further increasing simulation speed. The virtual prototype of the European FP7 project EURETILE system simulator is used as demonstrator for the proposed work, showing a speedup of 4.01x on a four core host system compared to sequential simulation.

[200] A Unified Methodology for a Fast Benchmarking of Parallel Architecture
Alexandre Guerre, Jean-Thomas Asquaviva, Yves Lhuillier

Benchmarking of architectures is today jeopardized by the explosion of parallel architectures and the dispersion of parallel programming models. Parallel programming requires architecture dependent compilers and languages as well as high programming expertise. Thus, an objective comparison has become a harder task. This paper presents a novel methodology to evaluate and to compare parallel architectures in order to ease the programmer work. It is based on the usage of micro-benchmarks, code profiling and characterization tools. The main contribution of this methodology is a semi-automatic prediction of the performance for sequential applications on a set of parallel architectures. In addition the performance estimation is correlated with the cost of other criteria such as power or portability. Our methodology prediction was validated on an industrial application. Results are within a range of 20%.
8.1 System Simulation and Virtual Prototyping

Organizer: Johannes Stahl, Synopsys, US; Chair: Johannes Stahl, Synopsys, US

Index Terms - System-level design, MP-SoC, ASP-algorithm joint optimization, multi-core architecture exploration

System-level Design Methodology Enabling Fast Development of Baseband MP-SoC for 4G Small Cell Base Station
Bernhard Fischer, Christian Cech, Hannes Muhr

Low power consumption of electronic devices has been an important requirement for many cyber-physical systems in field. Today, power dissipation is often estimated by spreadsheet-based power analysis. A leading-edge high-level power analysis method has the objective of providing high confidence levels in early design stages, where power design decisions have severe impact. This work examines and compares three high-level power analysis approaches (spreadsheet-based, Synopsys Platform Architect MCO, and DOcea Acceptor) by an industrial use case. The first chapter introduces general power analysis concepts. Chapter two presents different power analysis methods and tools that are applied on an industrial signal-processing use case described in chapter three. Chapter four compares these tools and methods by selected criteria such as power modeling effort and quality of results. Chapter five concludes about the investigated methods.

Keywords - power analysis; power modeling; early design phase; electronic system level; comparison; industrial use case

System-level Design Methodology Enabling Fast Development of Baseband MP-SoC for 4G Small Cell Base Station
Shan Tang, Ziyuan Zhu, Yongtao Su

"Small Cell" is regarded as the solution to optimize 4G wireless networks with improved coverage and capacity and expected to be deploy in a large number. To meet performance requirements and special constraints on the cost and size, we design a heterogeneous multi-processor SoC for small cell base station, which is composed of ASP (Application Specific Processor) cores, hardware accelerators, general-purpose processor core, and infrastructure and interface blocks. The challenges of developing such a complex chip drive us to employ system-level design methodology in both single core and multicores architecture optimizations. The paper discusses in detail the LISA (Language for Instruction-Set Architectures)/SystemC based ASP-algorithm joint optimization, and task-graph driven multi-core architecture exploration. Finally, the results of silicon implementation on SMIC 55nm technology are presented.

Index Terms - System-level design, MP-SoC, ASP-algorithm joint optimization, multi-core architecture exploration

Virtual Prototype Life Cycle in Automotive Applications
Manfred Thanner

Virtual prototypes for automotive applications see a unique life cycle in the context of the supply chain from semiconductor to Tier1 to OEMs and within the eco-system. The presentation gives an overview of current experiences and findings in this field and challenges observed. The virtual platforms targeting the mid to high end application spaces of chassis, to powertrain and driver information systems. The use cases primarily address today semiconductor internal developments and Tier1 level deployment. Additionally different software vendors use the models in their development cycle which drive model requirements like stimulus and abstraction levels. The development of virtual prototypes often start with the reuse of existing cores, accelerators and IP models. These models had certain use cases to address and were created accordingly. Therefore the models sometimes don‘t necessarily match fully the requirements
of the overall virtual prototype and compromises were made. Further to this, models are often from different design centers, vendors, etc. This can lead to conflicting model features versus the primary use case requirements of the virtual platform for the intended usage. Examples are cycle accuracy vs. functional, correct behavior vs. error behavior and error injection.

Keywords - Virtual Platform; System Level; SystemC; Automotive

8.2 Hot Topic: Near Threshold Computing (NTC)
Organizer: Michael Huebner, Ruhr-University Bochum, DE
Chair: Michael Huebner, Ruhr-University Bochum, DE

[208] Extreme-Scale Computer Architecture: Energy Efficiency from the Ground up
Josep Tormellas

As we move to integration levels of 1,000-core processor chips, it is clear that energy and power consumption are the most formidable obstacles. To construct such a chip, we need to rethink the whole compute stack from the ground up for energy efficiency and attain Extreme-Scale Computing. First of all, we want to operate at low voltage, since this is the point of maximum energy efficiency. Unfortunately, in such an environment, we have to tackle substantial process variation. Hence, it is important to design efficient voltage regulation, so that each region of the chip can operate at the most efficient voltage and frequency point. At the architecture level, we require simple cores organized in a hierarchy of clusters. Moreover, we also need techniques to reduce the leakage of on-chip memories and to lower the voltage guardbands of logic. Finally, data movement should be minimized, through both hardware and software techniques. With a systematic approach that cuts across multiple layers of the computing stack, we can deliver the required energy efficiencies.

[209] Voltage Island Management in Near Threshold Manycore Architectures to Mitigate Dark Silicon
Cristina Silvano, Gianluca Paletto, Sotiris Xydis, Ioannis Stamoulis

The power-wall problem driven by the stagnation of supply voltages in deep-submicron technology nodes, is now the major scaling barrier for moving towards the manycore era. Although the technology scaling enables extreme volumes of computational power, power budget violations will permit only a limited portion to be actually exploited, leading to the so-called dark silicon. Near-Threshold voltage Computing (NTC) has emerged as a promising approach to overcome the manycore power-wall, at the expenses of reduced performance values and higher sensitivity to process variations. Given that several application domains operate over specific performance constraints, the performance sustainability is considered a major issue for the widespread adoption of NTC. Thus, in this paper, we investigate how performance guarantees can be ensured when moving towards NTC manycores through variability-aware voltage and frequency allocation schemes. We propose three aggressive NTC voltage tuning and allocation strategies, showing that STC performance can be efficiently sustained or even optimized at the NTC regime. Finally, we show that NTC highly depends on the underlying workload characteristics, delivering average power gains of 65% for thread-parallel workloads and up to 90% for process-parallel workloads, while offering an extensive analysis on the effects of different voltage tuning/allocation strategies and voltage regulator configurations.

[210] Resolving the Memory Bottleneck for Single Supply Near-Threshold Computing
Tobias Gemmele, Mohammaded M. Saby, Jan Stuijt, Praveen Raghavan, Fransky Cattoor, David Atienza

This paper focuses on a review of state-of-the-art memory designs and new design methods for near-threshold computing (NTC). In particular, it presents new designs to reallocate low-voltage NTC memories cost-effectively by using reusable cell libraries, or by adding a digital wrapper around existing commercially available memories. The approach is based on modeling at system level supported by silicon measurement on a testchip in a 40nm low-power processing technology. Advanced monitoring, control and run-time error mitigation schemes enable the operation of these memories at the same optimal near-Vt voltage level as the digital logic. Reliability degradation is thus overcome and this opens the way to solve the memory bottleneck in NTC systems. Starting from the available 40 nm silicon measurements, the analysis is extended to future 14 and 10 nm technology nodes.

Keywords - Near-threshold computing; memories;
8.4 Efficient Designs for Telecom and Financial Applications

Chair: Sergio Saponara, University of Pisa, IT; Amer Baghdadi, Telecom Bretagne, FR

Energy Efficient MIMO: A Case Study of Opportunistic Run-Time Approximations
David Novo, Nazanin Farshooh, Paul Lee, Ubaid Ahmad, Franky Catthoor

Worst-case design is one of the keys to practical engineering: create solutions that can withstand the most adverse possible conditions. Yet, the ever-growing need for higher energy efficiency suggest a grim outlook for worst-case design in the future. In this paper, we propose opportunistic runtime approximations to enable a continuous adaptation of the processing precision (operator type and bitwidth) to the actual execution context without modifying the algorithm functionality. We show that by relaying the processing precision whenever possible, a VLSI implementation of an advanced wideband receiver algorithm based on opportunistic run-time approximations can save about 40% of the energy consumed by an optimized static implementation. These energy savings are achieved at the expense of a slight increase in overall chip area.

Energy-Efficient FPGA Implementation for Binomial Option Pricing Using OpenCL
Valentin Menas Morales, Pierre-Henri Homain, Amer Baghdadi, Erik Hochapfel, Sandrine Vaton

Energy efficiency of financial computations is a performance criterion that can no longer be dismissed, and is as crucial as raw acceleration and accuracy of the solution. In order to reduce the energy consumption of financial accelerators, FPGAs offer a good compromise with low power consumption and high parallelism. However, designing and prototyping an application on an FPGA-based platform are typically very time-consuming and requires significant skills in hardware design. This issue constitutes a major drawback with respect to software-centric acceleration platforms and approaches. A high-level approach has been chosen, using Altera’s implementation of the OpenCL standard, to answer this issue. We present two FPGA implementations of the binomial option pricing model on American options. The results obtained on a Terasic DE4 - Stratix IV board form a solid basis to hold all the constraints necessary for a real world application. The best implementation can evaluate more than 2000 options/s with an average power of less than 20W.

Hardware Implementation of a Reed-Solomon Soft Decoder Based on Information Set Decoding
Stefan Scholl, Norbert Wehn

Soft decision decoding of Reed-Solomon codes can largely improve frame error rates over currently used hard decision decoding. In this paper, we present a new hardware implementation for soft decoding of Reed-Solomon codes based on information set decoding. To our best knowledge this is the first hardware implementation of information set decoding for long Reed-Solomon codes. We propose a reduced complexity version of the decoding algorithm, that is optimized for efficient hardware implementation and enables high throughput.

The decoder was implemented on a Vortex 7 FPGA, achieving a gain of 0.75 dB compared to conventional hard decision decoding and a throughput of up to 1.19 GBit/s for the widely used RS(255,239). This gain in FER is achieved with less complexity and more than 15x larger throughput than other state-of-the-art architectures.

Ambient Variation-tolerant and Inter Components Aware Thermal Management for Mobile System on Chips
Francesco Patema, Joe Zanotelli, Tajana Simunic Rosing

In this work we measure and study two key aspects of the thermal behavior of smartphones: 1) thermal interaction between the components on the printed circuit board and 2) the influence of phone’s ambient temperature which is subject to large variations. The measurements on the smartphone running typical workloads show that the heat generated by the communication subsystem and the high temperatures on the back cover of the phone can increase the SoC temperature by as much as 17°C. None of the run-time thermal management studies presented to date considered this interaction, as there was no model available. We design a thermal model that captures this thermal dependency and a policy able to avoid thermal emergencies while minimizing the impact on performance.

Energy Efficient Data Flow Transformation for Givens Rotation Based QR Decomposition
Namita Sharma, Preeti Ranjan Panda, Min Li, Prashant Agrawal, Franky Catthoor

QR Decomposition (QRD) is a typical matrix decomposition algorithm that shares many common features with other algorithms such as LU and Cholesky decomposition. The principle can be realized in a large number of processing sequences that differ significantly in the number of memory accesses and computations, and hence, the overall implementation energy. With modern low power embedded processors evolving towards register file well connected with wide memory interfaces and vector functional units (FUs), the data flow in matrix decomposition algorithms needs to be carefully devised to ensure efficient energy implementation. In this paper, we present an efficient data flow transformation strategy for the Givens Rotation based QRD that optimizes data flow and memory accesses. We also explore different possible implementations for QRD of multiple matrices using the Simd feature of the processor. With the proposed data flow transformation, a reduction of up to 36% is achieved in the overall energy over conventional QRD sequences.

Mode-Controlled Dataflow Based Modeling & Analysis of a 4G-LTE Receiver
Hrishikesh Salunke, Orlando Moreira, Kees van Berkel

Today’s smartphones and tablets contain multiple cellular modems to support 2G/3G/4G standards, including Long Term Evolution (LTE). They run on complex multi-processor hardware platforms and have to meet hard real-time constraints. Dataflow modeling can be used to design an LTE receiver. Static dataflow allows a rich set of analysis techniques, but is too restrictive to model the dynamic behavior in many realistic applications, including LTE receivers. Dynamic dataflow allows modeling of many realistic applications, but does not support rigorous temporal analysis. Mode-Controlled Dataflow (MCDF) is a restricted form of dynamic dataflow, and allows the same analysis techniques as static dataflow, in principle. We prove that MCDF is sufficiently expressive to handle the dynamic behavior of a realistic LTE receiver, by systematically and stepwise developing a complete MCDF model for an LTE receiver.

8.5 Modeling & Specification

Chair: Wolfgang Mueller, University of Paderborn, DE; Francois PECHEX, UPAC, FR

An Activity-Sensitive Contention Delay Model for Highly Efficient Deterministic Full-System Simulations
Shu-Yung Chen, Chen-Hao Chen, Ren-Song Tsay

As modern systems are integrating exceeding number of components for better performance and functionality, early full-system simulation tools have become essential for validating complex concurrent system interaction activities. In the past decades, many useful timing-accurate system simulation tools have been developed; however, we find that even for the most efficient techniques, more than 90% of overhead occurs when simulating shared devices, such as buses. Instead of adopting the constant-delay model that compromises accuracy or using the time-consuming precise scheduling approach, we propose in this paper an effective system activity-sensitive contention delay model that can dynamically capture runtime contention situations and system configuration changes. To verify the idea, we construct an analytical bus delay model and integrate that into a system simulation tool. The experimental results show 20 to 80 times performance improvement over the scheduling-based bus model on full-system simulations and the estimated timing difference is less than 3%.

Automatic Specification Granularity Tuning for Design Space Exploration
Jiaxing Zhang, Gunar Schimmer

Algorithm Design Environments (ADE), such as Simulink, have been shown to be efficient for development, analysis, and evaluation of algorithms. Recent tools propose to facilitate algorithm architecture co-design by bridging the gap from ADE to System-Level Design Environments (SLDE) through automatic synthesis from algorithm models to SLDE specifications. With the wide range of block characteristic (from simple logic functions to complex kernels) in the algorithm model, however, it is challenging to select a suitable compositional granularity for SLDE Language (SLDL) blocks in the synthesized specification. A high volume of SLDL blocks of little computation will increase the number of mapping possibilities, whereas large blocks with heavy computation on the other hand allow inter-block fusion reducing the computational demands in the overall specification yet sacrificing the mapping flexibility. In this paper, we...
introduce an automatic specification granularity tuning mechanism to determine the granularity in the synthesized specification model hierarchy guided by the computational demands of algorithm blocks. Our granularity selection significantly simplifies the early design space exploration as only a meaningful block decomposition is exposed in the synthesized specification. It leads to an overall system with less computational demands by leveraging the block fusion capabilities in the ADE. At the same time our granularity decision ensures that sufficient flexibility remains in the system for exploring heterogeneous mapping of the algorithm. Our results on real-world examples show that specification models can be synthesized with 80% efficiency through block fusion with 70-90% fewer but coarser grained blocks.

(223) EDT: A Specification Notation for Reactive Systems
R. Venkatesh, Ulka Shrotri, G. Mural Krishna, Supriya Agrawal

Requirements of reactive systems express the relationship between sensors and actuators and are usually described in a natural language and a mix of state-based and stream-based paradigms. Translating these into a formal language is an important pre-requisite to automate the verification of requirements. The analysis effort required for the translation is a prime hurdle to formalization gaining acceptance among software engineers and testers. We present Expressive Decision Tables (EDT), a novel formal notation designed to reduce the translation efforts from both state-based and stream-based informal requirements. We have also built a tool, EDTTool, to generate test data and expected output from EDT specifications. In a case study consisting of more than 200 informal requirements of a real-life automotive application, translation of the informal requirements into EDT needed 43% lesser time than their translation into Statecharts. Further, we tested the Statecharts using test data generated by EDTTool from the corresponding EDT specifications. This testing detected one bug in a mature feature and exposed several missing requirements in another. The paper presents the EDTTool, notation, comparison to other notations and the details of the case study.

(224) Model-Based Actor Multiplexing with Application to Complex Communication Protocols
Christian Zebelein, Christian Haubelt, Joachim Falk, Tobias Schwarzer, Jürgen Teich

We propose a dynamic scheduling approach for the concurrent execution of logical actor instances on a single synthesized actor instance. Based on a formal dataflow model of computation, the proposed approach can be applied to a wide range of applications in a model-based design flow. As case study, we evaluate a bus-cycle-accurate SystemC RTL model based on an InfiniBand network adapter in a PCI Express system.

(225) A Novel Model for System-Level Decision Making with Combined ASP and SMT Solving
Alexander Bleuer, Jens Gludgau, Christian Haubelt

In this paper, we present a novel model enabling system-level decision making for time-triggered many-core architectures in automotive systems. The proposed application model includes shared data entities that need to be bound to memories during decision making. As a key enabler to our approach, we explicitly separate computation and shared memory communication over a network-on-chip (NoC). To deal with contention on a NoC, we model the necessary basis to implement a time-triggered scheduler that guarantees freedom of interference. We compute fundamental design decisions, namely (a) spatial binding, (b) multi-hop routing, and (c) time-triggered scheduling, by a novel coupling of answer set programming (ASP) with satisfiability modulo theories (SMT) solvers. First results of an automotive case study demonstrate the applicability of our method for complex real-world applications.

(226) DeSpErate: Speeding-up Design Space Exploration by Using Predictive Simulation Scheduling
Giovanni Manani, Gianluca Palermo, Vittorio Zaccaria, Chisna Silvano

The design space exploration (DSE) phase is used to tune configurable system parameters and it generally consists of a multiobjective optimization (MOO) problem. It is usually done at pre-design phase and consists of the evaluation of large design spaces where each configuration requires long simulation. Several heuristic techniques have been proposed in the past and the recent trend is reducing the exploration time by using analytic prediction models to approximate the system metrics, effectively pruning sub-optimal configurations from the exploration scope. However, there is still a missing path towards the effective usage of the underlying computing resources used by the DSE process. In this work, we will show that an alternative and almost orthogonal approach - focused on exploiting the available parallelism in terms of computing resources - can be used to better schedule the simulations and to obtain a high speedup with respect to state of the art approaches, without compromising the accuracy of exploration results. Experimental results will be presented by dealing with the DSE problem of a shared memory multi-core system considering a variable number of available parallel resources to support the DSE phase.

8.6 Mapping and Scheduling for Many-Core Embedded Systems

Chair: Marc Geelen, Eindhoven University of Technology, NL; Sébastien Le Beux, Ecole Centrale de Lyon, FR

(227) Software Architecture of High Efficiency Video Coding for Many-Core Systems with Power-Efficient Workload Balancing
Muhammad Usman Karim Khan, Muhammad Shafique, Jörg Henkel

The High Efficiency Video Coding (HEVC) standard aims at providing ~50% better compression compared to its predecessor (H.264) at the cost of high computational complexity. To enable HEVC video encoding in real-time scenarios, special coding support for parallelization is provided in HEVC that can be exploited by many-core systems. In this work, we present a HEVC software architecture where a video frame is adaptively divided into independent video frame regions (i.e. so-called video tiles) which are processed concurrently on multiple cores. By balancing the workload of each video tile mapped to a particular core, the total power consumption of a system is reduced through dynamically scaling the operating frequency under a given frame-rate constraint. We also exploit user tolerance to further curtail the HEVC workload with insignificant video quality degradation. Experimental results illustrate that the proposed approach results in ~43% power savings on a many-core system.

(228) GPU-EvR: Run-time Event Based Real-time Scheduling Framework on GPGPU Platform
Haseeung Lee, Muhammad Abdulrah Al Fakache

GPGPU architecture has traditionally been used in graphics application because of its enormous computing capability. Moreover, GPGPU architecture has also been used for general purpose computing in these days. Most of the current scheduling frameworks that are developed to handle GPGPU workload operate sequentially. This is problematic since this sequential approach may not be scalable for real-time systems, which is a consequence of the approach's inability to support preemption. We propose a novel scheduling framework that provides real-time support for the GPGPU platform. In contrast to existing frameworks, our proposed framework considers both concurrent execution of applications on the GPU and mapping between streaming multiprocessors and thread blocks. By considering both concurrent execution and mapping, our framework is able to guarantee timing up to 6.4 times as many applications compared to TimeGraph [9] and Global EDF [5]. In addition, our experimental applications use up to 25% less power under our scheduling framework compared to [5], [9].

(229) Multi-Objective Distributed Run-time Resource Management for Many-Cores
Stefan Wielandmann, Michael Glaj, Jürgen Teich

Dynamic usage scenarios of many-core systems require sophisticated run-time resource management that can deal with multiple often conflicting application and system objectives. This paper proposes an approach based on nonlinear programming techniques that is able to trade off between objectives while respecting targets regarding their values. We propose a distributed application embedding for dealing with soft system-wide constraints as well as a centralized one for strict constraints. The experiments show that both approaches may significantly outperform related heuristics.

(230) CoMik: A Predictable and Cycle-Accurately Composable Real-Time Microkernel
Andrew Nelson, Ashkan Beiranvand Nejad, Anca Molnos, Maritn Koedam, Kees Goossens

The functionality of embedded systems is ever increasing. This has lead to mixed time-critical systems, where applications with a variety of real-time requirements co-exist on the same platform and share resources. Due to inter-application interference, verifying the real-time requirements of such systems is generally non trivial. In this paper, we present the CoMik microkernel that provides temporally predictable and composable processor virtualisation. CoMik's virtual processors are cycle-accurately composable, i.e. their timing cannot affect the
8.7 Performance Modeling and Delay Test

Chairs: Robert Aitken, ARM, US; Mehdi Tahoori, KIT, DE

Myungsun Kim, Kibum Kim, James R. Geroa, Seongsoo Hong

ARM's big.LITTLE architecture introduces the opportunity to optimize power consumption by selecting the core type most suitable for a level of processing demand. To take advantage of this new axis of optimization, we introduce processor utilization into the Linux kernel's load balancing algorithm. Our method improves the Linux kernel's ability to schedule tasks in an energy efficient manner without making it directly aware of the available core types. Experimental results show an energy consumption improvement over the standard Linux scheduler up to 11.35% with almost no reduction in performance.

[232] hvecDTM: Application-Driven Dynamic Thermal Management for High Efficiency Video Coding
Daniel Paolino, Mohammad Shafique, Hussam Amrouch, Altamiro Sisan, Jorg Henkel

This paper presents an application-driven algorithm for Dynamic Thermal Management (DTM) for the High Efficiency Video Coding (HEVC). For efficient design of such a DTM policy, we perform an offline thermal analysis of an HEVC encoder and demonstrate the impact of different video sequences and different coding configurations on the processor temperature. Our thermal analysis is leveraged to develop an efficient application-driven DTM policy that performs temperature-aware coding along with an application-driven control of DTM knobs (e.g., frequency scaling) in order to meet the temperature constraints while still providing high video quality (i.e. PSNR loss < 0.01dB). For accurate thermal analysis and evaluation, we deploy an infrared camera-based thermal measurement setup that, on the contrary to state-of-the-art setups, does not require adding any extra layer on top of the measured chip, thus allowing the camera to accurately capture the infrared emissions from the die.

Keywords: Dynamic Thermal Management, DTM, HEVC, Application-Specific Optimization, Thermal Analysis, IR Camera.

[233] Improving Efficiency of Extensible Processors by Using Approximate Custom Instructions
Mehdi Kamal, Amin Ghassemzadeh, Ali Atzaki-Kusha, Massoud Pedram

In this paper, we propose to move the conventional extensible processor design flow to the approximate computing domain to gain more speedup. In this domain, the instruction set architecture (ISA) design flow selects both exact and approximate custom instructions (CIs). The proposed approach could be used for the applications where imprecise results may be tolerated. In the CI identification phase of the flow, the CIs which do not satisfy the maximum propagation delay but can provide approximate results also may be included in the CI candidate set. Next, in the selection phase, we propose a merit function which selects CIs with higher cycle savings and small error rates. The efficacy of the proposed approximate design flow is investigated using the case studies of the discrete cosine transform (DCT) and inverse DCT (iDCT) of the MPEG2 application. Also, the impact of the process variation on the imprecision of the results is investigated.

8.7.1 Performance Modeling and Delay Test

Li Yu, Shashid Saxena, Christopher Hess, Ibrahim (Abe) M. Elrefai, Dimitri Antoniadis, Duane Boring

In this paper, we propose a novel integrated circuits performance estimation algorithm through a physical subspace projection and maximum-a-posteriori (MAP) estimation. Our goal is to estimate the distribution of a target circuit performance with very small measurement sample size from on-chip monitor circuits. The key idea in this work is to exploit the fact that simulation and measurement data are physically correlated under different circuit configurations and topologies. First, different groups of measurements are projected to a subspace spanned by a set of projected variables. The projection is achieved by performing a sensitivity analysis of measurement parameters with respect to the subspace variables using a virtual-source MOSFET compact model. Then, a Bayesian treatment is developed by introducing prior distributions over these subspace variables. Maximum a posteriori estimation is then applied using the prior, and an expectation-maximization (EM) algorithm is used to estimate the circuit performance. The proposed method is validated by postsilicon measurement for a commercial 28-nm process. An average error reduction of 2x is achieved which can be translated to 32x reduction on data size needed for samples on the same die. A 150x and 70x sample size reduction on training dies is also achieved compared to traditional least-square fitting method and least-angle regression method, respectively, without reducing accuracy.

[235] Joint Virtual Probe: Joint Exploration of Multiple Test Items' Spatial Patterns for Efficient Silicon Characterization and Test Prediction
Shaoguang Zhang, Fan Lin, Chun-Kai Hsu, Kwang-Ting Cheng, Hong Wang

Virtual Probe (VP), proposed for characterization of spatial variations and for test time reduction, can effectively reconstruct the spatial pattern of a test item for an entire wafer using measurement values from only a small fraction of dies on the wafer. However, VP calculates the spatial signature of each test item separately, one item at a time, resulting in very long runtime for complex chips which often require hundreds, or even thousands, of test items in production. In this paper, we propose a new method, named Joint Virtual Probe (JVP), which can jointly derive spatial patterns of multiple test items. By simultaneously handling a large group of test items, JVP significantly reduces the overall runtime. And the prediction accuracy can also be improved because of JVP's implicit use of inter-test-item correlations in predicting spatial patterns. The experimental results on two industrial products, with 277 and 985 parametric test items in the test production programs respectively, demonstrate that, JVP achieves an average speedup of ~170x and ~50x over VP in the pre-test analysis and the test application phases respectively, as well as a slightly higher prediction accuracy than VP.

[236] Substituting Transition Faults with Path Delay Faults as a Basic Delay Fault Model
Irith Pomeranz

Comparing a single transition fault with a single path delay fault, targeting (i.e., simulating or generating a test for) a path delay fault is not more complex than targeting a transition fault. However, targeting a set of path delay faults is significantly more complex than targeting a set of transition faults when the goal is to consider the testable path delay faults that are associated with the longest paths. The reason is the large fraction of untestable path delay faults among these faults. This complication is removed if the requirement on the lengths of the paths is removed. In this case, it is possible to use path delay faults instead of transition faults as a basic delay fault model for better coverage of small delay defects. This paper studies the effects of using path delay faults as a basic delay fault model instead of transition faults.

[237] Standard Cell Library Tuning for Variability Tolerant Designs
Sebastian Fabian, Juan Diego Echeveria, Maarten Vertregt, José Pineda de Gyvez

In today's semiconductor industry we see a move towards smaller technology feature sizes. These smaller feature sizes pose a problem due to mismatch between identical cells on a single die known as local variation. In this paper a library tuning method is proposed which makes a smart selection of cells in a standard cell library to reduce the design's sensitivity to local variability. This results in a robust IC design with an identifiable behavior towards local variations. Experimental results performed on a widely used microprocessor design synthesized for a high performance timing show that we can achieve a timing spread reduction of 37% at an area increase cost of 7%.

[238] Probabilistic Standard Cell Modeling Considering Non-Gaussian Parameters and Correlations
André Lange, Christoph Schramm, Roland Jancke, Joachim Haase, Ingolf Lorenz, Ulf Schlichtmann

Variability continues to pose challenges to integrated circuit design. With statistical static timing analysis and high-yield estimation methods, solutions to particular problems exist, but they do not allow a common view on performance variability including potentially correlated and non-Gaussian parameter distributions. In this paper, we present a probabilistic approach for variability modeling as an alternative: model parameters are treated as multi-dimensional random variables. Such a fully multi-variate statistical description formally accounts for correlations
and non-Gaussian random components. Statistical characterization and model application are introduced for standard cells and gate-level digital circuits. Example analyses of circuitry in a 28 nm industrial technology illustrate the capabilities of our modeling approach.

8.8 Hot Topic: Beyond CMOS Ultra-low-power Computing
Organizer: Saibal Mukhopadhyay, Georgia Institute of Technology, US
Chairs: Arjil Rayachotthry, Georgia Institute of Technology, US; Saibal Mukhopadhyay, Georgia Institute of Technology, US

Ultra-low Power Electronics with SiGe Tunnel FET
Amit Rangan Trivedi, Mohammad Faisal Arif, Saibal Mukhopadhyay

SiGe Tunnel FET (TFET) with its subthreshold subthreshold swing is attractive for low power analog and digital designs. Greater Ion/Ioff ratio of TFET can reduce the dynamic power in digital designs, while higher gm/IDs can lower the bias power of analog amplifier. However, the above benefits of TFET are eclipsed by MOSFET at a higher power/performace point. Ultra low power scalability of the key analog and digital circuits, SRAM and operational transconductance amplifier (OTA), with TFET is demonstrated. Analyzing a TFET based cellular neural network, this work shows the feasibility of ultra-low-power neuromorphic computing with TFET.

Keywords - Tunnel FET, Low power design, SRAM, Operational transconductance amplifier, Cellular neural network

Brain-Inspired Computing with Spin Torque Devices
Kaushik Ray, Mihirak Sharad, Deliang Fan, Karthik Yogendra

In this paper we discuss the potential of emerging spintronic devices for computing applications. Recent proposals for spin-based computing schemes may be differentiated as 'all-spin' vs. hybrid, programmable vs. fixed, and, Boolean vs. non-Boolean. Allspin logic-styles may offer high area-density due to small form-factor of nano-magnetic devices. However, circuit and system-level design techniques need to be explored that leverage the specific spin-device characteristics to achieve energy-efficiency, performance and reliability comparable to those of CMOS. The non-volatility of nanomagnets can be exploited in the design of energy and area-efficient programmable logic. In such logic-styles, spin-devices may play the dual-role of computing as well as memory-elements that provide field-programmability. Spin-based threshold logic design is presented as an example. Emerging spintronic phenomena may lead to ultralow voltage, current-mode, spin-torque switches that can offer attractive computing capabilities, beyond digital switches. Such devices may be suitable for non-Boolean data-processing applications which involve analog processing. Integration of such spin-torque devices with charge-based devices like CMOS and resistive memory can lead to highly energy-efficient information processing hardware for applications like pattern-matching, neuromorphic computing, image processing and data-conversion. Finally, we discuss the possibility of using coupled spin-torque nano oscillators for lowpower non-Boolean computing.

Keywords: spin, logic, low power, threshold logic, analog, neural networks, non-Boolean, programmable logic array, coupled spin torque nano oscillators

III-V Semiconductor Nanowires for Future Devices
H. Schmidt, B.M. Bogg, K. Moselund, P. Das Kunungo, G. Signorello, S. Karg, P. Mensch, V. Schmidt, H. Riel

The monolithic integration of III-V nanowires on silicon by direct epitaxial growth enables new possibilities for the design and fabrication of electronic as well as optoelectronic devices. We demonstrate a new growth technique to directly integrate III-V semiconducting nanowires on silicon using selective area epitaxy within a nanobipolar transistor. Thus we achieve small diameter nanowires, controlled doping profiles and sharp heterojunctions crucial for future device applications. We experimentally demonstrate vertical tunnel diodes and gate-all-around tunnel FETs based on InAs/InP heterojunctions. The results indicate the benefits of the InAs/GaAs material system combining the possibility of achieving high Ion with high Ion/Ioff ratio.

Keywords - Tunnel FETs, Esaki diodes, nanowires, III-V semiconductors

Advanced System on a Chip Design Based on Controllable-Polarity FETs
Pierre-Emmanuel Guillardon, Luca Amaru, Jian Zhang, Giovanni De Micheli

Field-Effect Transistors (FETs) with on-off controllable-polarity are promising candidates to support next generation System-on-Chip (SoC). Thanks to their enhanced functionality, controllable-polarity FETs enable a superior design of critical components in a SoC such as processing units and memories, while also providing native solutions to control power consumption. In this paper, we present the efficient design of a SoC core with controllable-polarity FET. Processing units are speeded-up at the datapath level, as arithmetic operations require fewer physical resources than in standard CMOS. Power consumption is decreased via embedded-polarizing gates and tunable high-performance low-power devices. Memory cells are made smaller by merging the access interface with the storage circuitry. We foresee the advantages deriving from these techniques, by evaluating their impact on the design of SoC for a contemporary telecommunication application. Using a 22-nm vertically-stacked silicon nanowire technology, a coarse-grain evaluation at the block level estimates a delay and power reduction of 20% and 18% respectively, at a cost of a moderate area overhead of 15%, with respect to a state-of-art FinFET technology.

Keywords - Functionality-enhanced devices; System-on-Chip; Datapath; Low-power techniques

Reconfigurable Silicon Nanowire Devices and Circuits: Opportunities and Challenges
Walter M. Weber, Jens Tommer, Matthias Grube, André Heinzig, Markus König, Thomas Mikolajczyk

Reconfigurable fine-grain electronics target an increase in the number of integrated logic functions per chip by enhancing the functionality at the device level and by implementing a compact and technologically simple hardware platform. Here we study a promising realization approach by employing reconfigurable nanowire transistors (RFETs) as the multifunctional building-blocks to be integrated therein. RFETs merge the electrical characteristics of unipolar n- and p-type FETs into a single universal device. The switch comprises four terminals, where three of them act as the conventional FET electrodes and the fourth acts as an electric select signal to dynamically program the desired switch type. The transistor consists of two independent charge carrier injection valves as represented by two gated Schottky junctions integrated within an intrinsic silicon nanowire. Radial compressive strain applied to the channel is used as a scalable method to adjust n- and p-FET currents to each other, thereby enabling complementary logic circuits. Simple but relevant examples for the reconfiguration of complete gates will be given, demonstrating the potential of this technology.

Keywords - Reconfigurable transistor, RFET, nanowire, Schottky FET, reconfigurable circuit, inverter, doping free CMOS, symmetric FET, universal transistor

9.1 Hot Topic: CMOS Scaling - from Evolutionary to Revolutionary Computing
Organizer: Thomas Mikolajczyk, NamLab gGmbH, DE; Ian O'Connor, Lyon Institute of Nanotechnology, FR
Chairs: Thomas Mikolajczyk, NamLab gGmbH, DE; Ian O'Connor, Lyon Institute of Nanotechnology, FR

Advanced System on a Chip Design Based on Controllable-Polarity FETs
Pierre-Emmanuel Guillardon, Luca Amaru, Jian Zhang, Giovanni De Micheli

Field-Effect Transistors (FETs) with on-off controllable-polarity are promising candidates to support next generation System-on-Chip (SoC). Thanks to their enhanced functionality, controllable-polarity FETs enable a superior design of critical components in a SoC such as processing units and memories, while also providing native solutions to control power consumption. In this paper, we present the efficient design of a SoC core with controllable-polarity FET. Processing units are speeded-up at the datapath level, as arithmetic operations require fewer physical resources than in standard CMOS. Power consumption is decreased via embedded-polarizing gates and tunable high-performance low-power devices. Memory cells are made smaller by merging the access interface with the storage circuitry. We foresee the advantages deriving from these techniques, by evaluating their impact on the design of SoC for a contemporary telecommunication application. Using a 22-nm vertically-stacked silicon nanowire technology, a coarse-grain evaluation at the block level estimates a delay and power reduction of 20% and 18% respectively, at a cost of a moderate area overhead of 15%, with respect to a state-of-art FinFET technology.

Keywords - Functionality-enhanced devices; System-on-Chip; Datapath; Low-power techniques

Reconfigurable Silicon Nanowire Devices and Circuits: Opportunities and Challenges
Walter M. Weber, Jens Tommer, Matthias Grube, André Heinzig, Markus König, Thomas Mikolajczyk

Reconfigurable fine-grain electronics target an increase in the number of integrated logic functions per chip by enhancing the functionality at the device level and by implementing a compact and technologically simple hardware platform. Here we study a promising realization approach by employing reconfigurable nanowire transistors (RFETs) as the multifunctional building-blocks to be integrated therein. RFETs merge the electrical characteristics of unipolar n- and p-type FETs into a single universal device. The switch comprises four terminals, where three of them act as the conventional FET electrodes and the fourth acts as an electric select signal to dynamically program the desired switch type. The transistor consists of two independent charge carrier injection valves as represented by two gated Schottky junctions integrated within an intrinsic silicon nanowire. Radial compressive strain applied to the channel is used as a scalable method to adjust n- and p-FET currents to each other, thereby enabling complementary logic circuits. Simple but relevant examples for the reconfiguration of complete gates will be given, demonstrating the potential of this technology.

Keywords - Reconfigurable transistor, RFET, nanowire, Schottky FET, reconfigurable circuit, inverter, doping free CMOS, symmetric FET, universal transistor
9.2 Low-Cost, High-Performance NoCs

Chairs: Kees Goossens, Eindhoven University, NL; Luca Ramini, University of Ferrara, IT

[246] Application Mapping for Express Channel-Based Networks-on-Chip
Di Zhu, Lichong Chen, Siyu Yue, Massoud Pedram

With the emergence of many-core multiprocessor system-on-chips (MPSoCs), the on-chip networks are facing serious challenges in providing fast communication for various tasks and cores. One promising solution shown in recent studies is to add express channels to the network as shortcuts to bypass intermediate routers, thereby reducing packet latency. However, this approach also greatly changes the packet delay estimation and traffic behaviors of the network, both of which have not yet been exploited in existing mapping algorithms. In this paper, we explore the opportunities in optimizing application mapping for express channel-based on-chip networks. Specifically, we derive a new delay model for this type of networks, identify their unique characteristics, and propose an efficient heuristic mapping algo-rithm that increases the bypassing opportunities by reducing unnecessary turns that would otherwise impose the entire router pipeline delay to packets. Simulation results show that the proposed algorithm can achieve a 2-4X reduction in the number of turns and 10-26% reduction in the average packet delay.

Keywords - network-on-chip; application mapping; express channels

[247] Parallel Probe Based Dynamic Connection Setup in TDM NoCs
Shaoteng Liu, Axel Janisch, Zhonghai Lu

We propose a Time-Division Multiplexing (TDM) based connection oriented NoC with a novel double time-wheel router architecture combined with a run-time parallel probing setup method. Compared with traditional TDM connection setup methods, our design has the following advantages: (1) it allocates paths and time slots at run-time; (2) it is fast with predictable and bounded setup latency; (3) it avoids additional resources (no auxiliary network or central processor to find and manage connections); (4) it is fully distributed and therefore it scales nicely with network size. Compared to a packet based setup method, our probe based design can reduce path setup delay by 81% and increase network load by 110% in an 8x8 mesh, while avoiding the auxiliary network. Compared to a centralized method, our solution can double the success rate, while eliminating the central resource for path setup and reducing the wire overhead. Synthesis results suggest that our design is faster and smaller than all comparable solutions.

I. Seitander, A. Piaans, G. Dimitrakopoulos, C. Nicopoulos

The design of scalable Network-on-Chip (NoC) architectures calls for new implementations that achieve high-throughput and low-latency operation, without exceeding the stringent area-energy constraints of modern Systems-on-Chip (SoC). The router's buffer architecture is a critical design aspect that affects both network-wide performance and implementation characteristics. In this paper, we extend Elastic Buffer (EB) architectures to support multiple Virtual Channels (VC) and we derive ElastiStore, a novel lightweight elastic buffer architecture that minimizes buffering requirements, without sacrificing performance. The integration of the proposed elastic buffering scheme in the NoC router enables the design of new router architectures - both single-cycle and two-stage pipelined - which offer the same performance as baseline VC-based routers, albeit at a significantly lower area/power cost.

[249] Dynamic Construction of Circuits for Reactive Traffic in Homogeneous CMPs
Martín Ortín, Dario Sudrez, María Villamayor, Cruz Izu, Victor Viala

Networks on Chip (NoC) have a large impact on system performance, area and energy. Considering the characteristics of the memory subsystem while designing the NoC helps identify improvement opportunities and build more efficient designs. Leveraging the frequent request-reply pattern, our proposal dynamically builds the reply path in advance, is able to share circuits between messages, and even removes some implicit replies, significantly reducing NoC latency. A careful implementation of this circuit reservation mechanism achieves an average 17% reduction in router energy consumption, 8% smaller router area, and a 2% system performance increase, compared with its baseline counterpart.

[250] Improving Hamiltonian-based Routing Methods for On-chip Networks: A Turn Model Approach
Poona Bahnebar, Deh Stroobandt

The overall performance of Multi-Processor System-on-Chip (MPSoC) platforms depends highly on the efficient communication among their cores in the Network-on-Chip (NoC). Routing algorithms are responsible for the on-chip communication and traffic distribution through the network. Hence, designing efficient and high-performance routing algorithms is of significant importance. In this paper, a deadlock-free and highly adaptive path-based routing method is proposed without using virtual channels. This method strives to exploit the maximum number of minimal paths between any source and destination pair. The simulation results in terms of performance and power consumption demonstrate that the proposed method significantly outperforms the other adaptive and non-adaptive schemes. This efficiency is achieved by reducing the number of hotspots and smoothly distributing the traffic across the network.

9.3 Hardware Implementations for Data Security

Chairs: Viktor Fischer, St Etienne, FR; Filippo Melzani, STMicroelectronics, IT

[251] Embedded Reconfigurable Logic for ASIC Design Obfuscation against Supply Chain Attacks
Bao Liu, Brandon Wang

Hardware is the foundation and the root of trust of any security system. However, in today's global IC industry, an IP provider, an IC design house, a CAD company, or a foundry may subvert a VLSI system with back doors or logic bombs. Such a supply chain adversary's capability is rooted in his knowledge on the hardware design. Successful hardware design obfuscation would severely limit a supply chain adversary's capability if not preventing all supply chain attacks. However, not all designs are obfuscatable in traditional technologies. We propose to achieve ASIC design obfuscation based on embedded reconfigurable logic which is determined by the end user and unknown to any party in the supply chain. Combined with other security techniques, embedded reconfigurable logic can provide the root of ASIC design obfuscation, data confidentiality and tamper-proofness. As a case study, we evaluate hardware-based code injection attacks and reconfiguration-based instruction set obfuscation based on an open source SPARC processor LEON2. We prevent program monitor Trojan attacks and increase the area of a minimum code injection Trojan with a 1KB ROM by 2.38X for every 1% area increase of the LEON2 processor.

[252] A Minimalist Approach to Remote Attestation
Aurélien Francillon, Quan Nguyen, Kasper B. Rasmussen, Gene Tsudik

Advancing CMOS with Carbon Electronics
Franz Kreupl

A fresh look on carbon-based transistor channel materials like single-walled carbon nanotubes (CNT) and graphene nanoribbons (GNR) in future electronic applications is given. Although theoretical predictions initially promised that GNR (which do have a bandgap) would perform equally well as transistors based on CNTs, experimental evidence for the well-behaved transistor action is missing up to now. Possible reasons for the shortcomings as well as possible solutions to overcome the performance gap will be addressed. In contrast to GNR, short channel CNT field effect transistors (FET) demonstrate in the experimental realization almost ideal transistor characteristics down to very low bias voltages. Therefore, CNT-FETs are clear frontrunners in the search of a future CMOS switch, that will enable further voltage and gate length scaling. Essential features which distinguish CNT-FETs from alternative solution will be discussed and benchmarked. Finally, the gap to industrial water-level scale SWCNT integration will be addressed and strategies for achieving highly aligned carbon nanotube fabrics will be discussed. Without such a high yield wafer-scale integration, SWCNT circuits will be an illusional dream.

Keywords - carbon nanotube; graphene; nanoribbon; electronic; transistor; integration
Embedded computing devices increasingly permeate many aspects of modern life: from medical to automotive, from building and factory automation to weapons, from critical infrastructures to home entertainment. Despite their specialized nature as well as limited resources and connectivity, these devices are now becoming an increasingly popular and attractive target for attacks, especially malware infections. A number of approaches have been suggested to detect and/or mitigate such attacks. They vary greatly in terms of application generality and underlying assumptions. However, one common theme is the need for Remote Attestation, a distinct security service that allows a trusted party (verifier) to check the internal state of a remote untrusted embedded device (prover). Many prior methods assume some form of trusted hardware on the prover, which is not a good option for small and low-end embedded devices. To this end, we investigate the feasibility of Remote Attestation without trusted hardware. This paper provides a systematic treatment of Remote Attestation, starting with a precise definition of the desired service and proceeding to its systematic deconstruction into necessary and sufficient properties. Next, these are mapped into a minimal collection of hardware and software components that result in secure Remote Attestation. One distinguishing feature of this line of research is the need to prove (or, at least argue for) architectural minimalism - an aspect rarely encountered in security research. This work also provides a promising platform for attaining more advanced security services and guarantees.

In today's technology driven world, it is essential to build secure systems with low faulty behavior. Authentication is one of the primary means to gain access to secure systems. Users need to be authenticated in order to gain access to the services and sensitive information contained within the system. Due to the surge in the number of touch based smart devices, there arises a need for a compatible authentication system. Historically, fingerprints have served in its fullest capacity to establish the uniqueness of an individual's identity. It can be detected using capacitive sensing techniques. In this paper we present a novel unified device using transparent electronics for both fingerprint scan and multi-touch interaction. We discuss a high resolution transparent touch sensitive device and a read out circuit that drives the capacitive sensor array for touch interactions at low resolutions and for fingerprint sensing at higher resolutions. Using circuit simulation and custom Verilog: A model for transparent thin-film transistors, we verified that our design can sense fingerprints in 8.25 ms and detect touches in 0.6ms with an efficient power consumption of 1 mW. The results show that such a device can be realized and can serve as a very efficient means of user authentication. Furthermore, from the usability aspect, the proposed device is essential as it provides user transparent and non intrusive authentication.

As cloud computing becomes mainstream, the need to ensure the privacy of the data entrusted to third parties keeps rising. Cloud providers resort to numerous security controls and encryption to thwart potential attackers. Still, since the actual computation inside cloud microprocessors remains unencrypted, the opportunity of leakage is theoretically possible. Therefore, in order to address the challenge of protecting the computation inside the microprocessor, we introduce a novel general purpose architecture for secure data processing, called HEROIC (Homomorphically Encrypted One Instruction Computer). This new design utilizes a single instruction architecture and provides native processing of encrypted data at the architecture level. The security of the solution is assured by a variant of Pailler's homomorphic encryption scheme, used to encrypt both instructions and data. Experimental results using our hardware-cognizant software simulator, indicate an average execution overhead between 5 and 45 times for the encrypted computation (depending on the security parameter), compared to the unencrypted variant, for a 16-bit single instruction architecture.

The security concerns of EDA tools have long been ignored because IC designers and integrators only focus on their functionality and performance. This is a lack of trusted EDA tools hampers hardware security researchers' efforts to design trusted integrated circuits. To address this concern, a novel EDA tools trust evaluation framework has been proposed to ensure the trustworthiness of EDA tools through its functional integration, rather than scrutinizing the software code. As a result, the newly proposed framework lowers the evaluation cost and is a better fit for hardware security researchers. To support the EDA tools evaluation framework, a new gate-level information assurance scheme is developed for security property checking on any gatelevel netlist. Helped by the gate-level scheme, we expand the territory of proof-carrying based IP protection from RT-level designs to gate-level netlist, so that most of the commercially trading third-party IP cores are under the protection of proof-carrying based security properties. Using a sample AES encryption core, we successfully prove the trustworthiness of Synopsys Design Compiler in generating a synthesized netlist.

9.4 Timing challenges in validation

Chair: Elena Ioana Vatajelu, Politecnico di Torino, IT; Mark Zwölfer, University of Southampton, UK

Fast STA Prediction-based Gate-level Timing Simulation
Tang B. Ahmad, Macej J. Ciesielski

Traditional dynamic simulation with standard delay format (SDF) back-annotation cannot be reliably performed on large designs. The large size of SDF files makes the event-driven simulation extremely slow as it has to process an excessive number of events. In order to accelerate gate-level timing simulation we propose an automated fast prediction-based gate-level timing simulation that combines static timing analysis (STA) at the block level with dynamic timing simulation at the I/O interfaces. We demonstrate that the proposed timing simulation can be done earlier in the design cycle in parallel with synthesis.

A Cross-Level Verification Methodology for Digital IPs Augmented with Embedded Timing Monitors
V. Guarnieri, M. Petricca, A. Sassone, S. Vinco, N. Bombieri, F. Fummi, E. Macii, M. Poncino

Smart systems implement the leading technology advances in the context of embedded devices. Current design methodologies are not suitable to deal with tightly interacting subsystems of different technological domains, namely analog, digital, discrete and power devices, MEMS and power sources. The effects of interaction between components and with the environment must be modeled and simulated at system level to achieve high performance. Focusing on the digital domain, additional design constraints have to be considered as a result of the integration of multi-domain subsystems in a single device. The main digital design challenges, common to those emerging from the heterogeneous nature of the whole system, directly impact on performance and on propagation delay of the digital component. This paper proposes a design approach to enhance the RTL model of a given digital component for the integration in smart systems, and a methodology to verify the added features at system-level. The design approach consists of augmenting the RTL model through the automatic insertion of delay sensors, which can detect and correct timing failures. The augmented model is abstracted to SystemC TLM and, then, mutants (i.e., code mutations for emulating timing failures) are automatically injected into the model. Experimental results demonstrate the applicability of the proposed design and verification methodology and the effectiveness of the simulation performance.

Empowering Study of Delay Bound Tightness with Simulated Annealing
Xueqian Zhao, Zhonghai Lu

Studying the delay bound tightness typically takes a practical approach by comparing simulated results against analytic results. However, this is often a manual process whereas many simulation parameters have to be configured before the simulations run. This is a tedious and time-consuming process. We propose a technique to automate this process by using a simulated annealing approach. We formulate the problem as an online optimization problem, and embed a simulated annealing algorithm in the simulation environment to guide the search of configuration parameters which give good tightness results. This is a fully automated procedure and thus provide a promising path to automatic design space exploration in similar contexts. Experimental results of an all-to-one communication network with large searching space and complicated constraints illustrate the effectiveness of our method.

Analysis and Evaluation of Per-Flow Delay Bound for Multiplexing Models
Yanchen Long, Zhonghai Lu, Xiaolang Yan
Multiplexing models are common in resource sharing communication media such as buses, crossbars and networks. While sending packets over a multiplexing node, the packet delay bound can be computed using network calculus models. The tightness of such delay bound remains an open problem. This paper studies the multiplexing models for weighted round robin scheduling with different traffic matrixes, and analyzes per-flow packet delay bounds with different service properties. We empirically evaluate the tightness of the delay bounds. Our results show the quality of different analysis models, and how influential each parameter is to tightness.

9.6 Schedulability Analysis
Chair: Rob Davis, University of York, UK; Giuseppe Lipari, ENS - Cachan, FR

General and Efficient Response Time Analysis for EDF Scheduling
Nan Guan, Wang Yi
Response Time Analysis (RTA) is one of the key problems in real-time system design. This paper proposes new RTA methods for EDF scheduling, with general system models where workload and resource availability are represented by request/command bound functions and supply bound functions. The main idea is to derive response time upper bounds by lower-bounding the slack times. We first present a simple over-approximate RTA method, which lower bounds the slack time by measuring the "horizontal distance" between the demand bound function and the supply bound function. Then we present an exact RTA method based on the above idea but eliminating the pessimism in the first analysis. This new exact RTA method, not only allows to precisely analyze more general system models than existing EDF RTA techniques, but also significantly improves analysis efficiency. Experiments are conducted to show efficiency improvement of our new RTA technique, and tradeoffs between the analysis precision and efficiency of the two methods in this paper are discussed.

The Schedulability Region of Two-Level Mixed-Criticality Systems Based on EDF-VD
Dirk Müller, Alejandro Masur
The algorithm Earliest Deadline First with Virtual Deadlines (EDF-VD) was recently proposed to schedule mixed-criticality task sets consisting of high-criticality (HI) and low-criticality (LO) tasks. EDF-VD distinguishes between HI and LO mode. In HI mode, the HI tasks may require executing for longer than in LO mode. As a result, in LO mode, EDFVD assigns virtual deadlines to HI tasks (i.e., it uniformly downscales deadlines of HI tasks) to account for an increase of workload in HI mode. Different schedulability conditions have been proposed in the literature; however, the schedulability region to fully characterize EDF-VD has not been investigated so far. In this paper, we review EDF-VD's schedulability criteria and determine its schedulability region to better understand and design mixed-criticality systems. Based on this result, we show that EDF-VD has a schedulability region being around 85% larger than that of the Worst-Case Reservations (WCR) approach.
Keywords: real-time scheduling; mixed-critically; EDF-VD; resource efficiency

9.7 Timing Analysis and Cell Design
Chair: Jose Monteiro, INESC-ID / Tecnoim, ULisboa, PT; Elena Dubrova, Royal Institute of Technology, SE

Facilitating Timing Debug by Logic Path Correspondence
Oshar Adler, Eli Ariel, Ilia Averbouch, Ilan Beer, Inna Grinievitch
Synthesis tools for high-performance VLSI designs employ aggressive logic optimization techniques in order to meet physical requirements such as area and cycle time. During these optimizations, the original structure of the design, which is usually written in a hardware description language (HDL), is lost. It is difficult, and often impossible, to relate signals after synthesis to the original signals in the HDL code. Some signals only lose their names while for others there are no equivalent counterparts in the design after synthesis. Debugging timing problems is based on timing reports which are usually represented in terms of the post-synthesis design. Hence, it is difficult to relate critical paths in the timing reports to the relevant paths in the HDL code when a logic fix is needed. In this paper, we propose a different approach for dealing with the correspondence problem: instead of trying to relate signals we relate paths. Given a critical path in a post-synthesis representation, our method is able to find all corresponding paths in the pre-synthesis (HDL) representation. As a result, locating the parts in the HDL which are relevant to the given timing problem becomes trivial. A novel SAT based algorithm for dealing with the path correspondence problem is described. Experimental results on various industrial high-end processor designs show the effectiveness of our algorithm in substantially reducing the amount of paths in the HDL which one will have to consider when debugging a given critical path.

Statistical Static Timing Analysis Using a Skew-Normal Canonical Delay Model
Vijay Kumar M, V Vasudevan
In its simplest form, a parameterized block-based statistical static timing analysis (SSTA) is performed by assuming that both gate delays and the arrival times at various nodes are...
Energy-efficiency has emerged as a major barrier to performance scalability for modern processors. We note that significant part of processor's energy requirement is contributed by applications on user configurable computer system architectures can be evaluated in high accuracy. In addition, MSim provides flexible interfaces with pervasive object-oriented design, characterizing the two types of ADBs called CADB (capacitor based ADB) and IADB (inverter based ADB) and show that the adjusted delays by IADB can be decremented, and show that cost buffer sizing has never been or not been primarily taken into account. To demonstrate how much overcoming the two limitations is effective in resolving the clock skew constraint, we recently [1-4, 6]. As such, the fascinating idea of memcomputing (i.e., use memory for computation purposes) has drawn wide attention from both academia and industry as the clock skew violation in some clock trees of multiple power modes can be resolved by applying buffer sizing together with using only a small number of IADB. 

In this work we introduce a new logic style for p-n junctions based digital graphene circuits: the pass-XNOR logic style. The latter enables the realization of compact, energy efficient circuits excepted lifetime. We find an optimal tradeoff between the increase in the size of the library and the lifetime improvement (timing margin reduction) by non-uniform extension of the library cells for various ranges of the input signal probabilities. The simulation results reveal that our technique can prolong the circuit lifetime by around 150% with a negligible area overhead.

Recently, many works have shown that adjustable delay buffer (ADB) whose delay is adjustable dynamically can effectively solve the clock skew variation problem in the designs with multiple power modes. However, all the previous works of ADB allocation inherently entail two critical limitations, which are the adjusted delays by ADB are always increments and the low cost buffer sizing has never been or not been primarily taken into account. To demonstrate how much overcoming the two limitations is effective in resolving the clock skew constraint, we characterize the two types of ADBs called CADB (capacitor based ADB) and IADB (inverter based ADB) and show that the adjusted delays by IADB can be decremented, and show that the clock skew violation in some clock trees of multiple power modes can be resolved by applying buffer sizing together with using only a small number of IADBs and CADBs.

In this paper, we propose a model based on skew-normal random variables. It can take into account the skewness in the gate delay distribution as well as the nonlinearity of the MAX function. We derive analytical expressions for the moments of the MAX operator based on the conditional expectations. The computational complexity of using this model is marginally higher than the linear model based on Clark's approximations. The results obtained using this model match well with Monte-Carlo simulations.

In the design of nonzero clock skew circuits, an increase of the path delay may improve circuit speed and reduce leakage power. However, the impact of increasing path delay on the trade-off between circuit speed and leakage power has not been well studied. In this paper, we propose a two-step approach for leakage-power-aware clock period minimization. Compared with previous works, our approach has the following two significant contributions: First, our approach is the first leakage-power-aware clock skew scheduling that can guarantee working with the lower bound of the clock period. Second, our approach is also the first work that demonstrates the problem of minimizing the number of extra buffers is a polynomial-time problem. Benchmark data show that our approach achieves the best results in terms of the clock period and the leakage power.

Keywords - Clock Period Minimization, Leakage Power, Clock Slew Scheduling, Sequential Timing Optimization.

A Deep Learning Methodology to Proliferate Golden Signoff Timing Seung-Soo Han, Andrew B. Kahng, Satishantha Nath, Ashok S. Vydyanathan

Signoff timing analysis remains a critical element in the IC design flow. Multiple signoff corners, libraries, design methodologies, and implementation flows make timing closure very complex at advanced technology nodes. Design teams often want to ensure that one tool’s timing reports are neither optimistic nor pessimistic with respect to another tool’s reports. The resulting "correlation" problem is highly complex because tool contains millions of lines of black-box and legacy code, licenses prevent any reverse-engineering of algorithms, and the nature of the problem is seemingly "unbounded" across possible designs, timing paths, and electrical parameters. In this work, we apply a "big-data" approach to the timing correlation problem. We develop a machine learning-based tool, Golden Timer eXtension (GTX), to correct divergence in flip-flop setup time, cell arc delay, wire delay, stage delay, and path slack at timing endpoints between times. We propose a methodology to apply GTX to two arbitrary times, and we evaluate scalability of GTX across multiple designs and foundry technologies / libraries, both with and without signal integrity analysis. Our experimental results show reduction in divergence between timing tools from 139.3ps to 21.1ps (i.e., 6.6x) in endpoint slack, and from 117ps to 23.8ps (4.9x reduction) in stage delay. We further demonstrate the incremental application of our methods so that models can be adapted to any outlier discrepancies when new designs are taped out in the same technology / library. Last, we demonstrate that GTX can also correlate timing reports between signoff and design implementation tools.

Aging-aware Standard Cell Library Design Saman Kiamehr, Farshad Feizoo, Mojtaba Ebrahimi, Mehdi B. Tahoori

Transistor aging, mostly due to Bias Temperature Instability (BTI), is one of the major unreliability sources at nanoscale technology nodes. BTI causes the circuit delay to increase and eventually leads to a decrease in the circuit lifetime. Typically, standard cells in the library are optimized according to the design time delay, however, due to the asymmetric effect of BTI, the rise and fall delays might become significantly imbalanced over the lifetime. In this paper, the BTI effect is mitigated by balancing the rise and fall delays of the standard cells at the expected lifetime. We find an optimal tradeoff between the increase in the size of the library and the lifetime improvement (timing margin reduction) by non-uniform extension of the library cells for various ranges of the input signal probabilities. The simulation results reveal that our technique can prolong the circuit lifetime by around 150% with a negligible area overhead.

Mixed Allocation of Adjustable Delay Buffers Combined with Buffer Sizing in Clock Tree Synthesis of Multiple Power Mode Designs Kiseo Park, GeounHo Kim, TaeWahn Kim

Recently, many works have shown that adjustable delay buffer (ADB) whose delay is adjustable dynamically can effectively solve the clock skew variation problem in the designs with multiple power modes. However, all the previous works of ADB allocation inherently entail two critical limitations, which are the adjusted delays by ADB are always increments and the low cost buffer sizing has never been or not been primarily taken into account. To demonstrate how much overcoming the two limitations is effective in resolving the clock skew constraint, we characterize the two types of ADBs called CADB (capacitor based ADB) and IADB (inverter based ADB) and show that the adjusted delays by IADB can be decremented, and show that the clock skew violation in some clock trees of multiple power modes can be resolved by applying buffer sizing together with using only a small number of IADBs and CADBs.

Embedded Tutorial: Memcomputing: The Cape of Good Hope

Organizers: Yiyu Shi, Missouri University of Science & Technology, US; Hung-Ming Chen, National Chiao Tung University, Taiwan Chairs: Yiyu Shi, Missouri University of Science & Technology, US; Hung-Ming Chen, National Chiao Tung University, Taiwan

Memcomputing: The Cape of Good Hope Yiyu Shi, Hung-Ming Chen

Energy efficiency has emerged as a major barrier to performance scalability for modern processors. On the other hand, significant breakthroughs have been achieved in memory technologies recently [1-4, 6]. As such, the fascinating idea of memcomputing (i.e., use memory for computation purposes) has drawn wide attention from both academia and industry as an effective remedy. Compared with conventional logic computing, memory array provides large set of parallel resources with high bandwidth, which can be configured to perform in-situ computing and information processing, leading to drastic reduction in processor-memory traffic. It will not only make computations more power- and speed-efficient, but also smarter. In addition, memcomputing provides flexible interfaces with pervasive object-oriented design, which makes it well suited as a good base platform for researchers to explore new memcomputing technologies.
processor-memory communication. To address the energy issue in processors, we propose a novel hardware accelerator framework that transforms high-density memory array into a configurable computing resource to accelerate variety of tasks - both compute- and data-intensive. It exploits the block-based architecture of nanoscale memory to create a spatially connected array of lightweight processors, each of which uses a memory block as its local memory. The proposed framework provides some unique advantages for hardware acceleration compared to conventional accelerators: 1) memory array provides large set of parallel resources with high bandwidth, which can be configured to perform computing in spatio-temporal manner leading to dramatic reduction in processor-memory traffic; 2) it brings the computing engine close to the data, thus drastically minimizing the von Neumann bottleneck; 3) finally, it exploits the advances in memory technologies and integration approaches e.g. 3D integration to achieve better technology scalability compared to alternative reconfigurable accelerator platforms. Simulation results for several data-intensive applications show that the proposed computing approach provides significant improvement in energy-efficiency compared to software while achieving significantly lower hardware overhead.

10.1 Hot Topic: Memories Today and Tomorrow
Organizers: Thomas Mikolajck, NamLab gGmbH, DE; Ian O'Connor, Lyon Institute of Nanotechnology, FR
Chairs: Ian O'Connor, Lyon Institute of Nanotechnology, FR; Thomas Mikolajck, NamLab gGmbH, DE

[275] Exploring the Limits of Phase Change Memories
Matthias Wüstig
Phase change materials are among the most promising compounds in information technology. They can be very rapidly switched between the amorphous and the crystalline state, indicative for peculiar crystallization behaviour. Phase change materials are already employed in rewritable optical data storage, where the pronounced difference of optical properties between the amorphous and crystalline state is used. This unconventional class of materials is also the basis of a storage concept to replace flash memory. This talk will discuss the unique material properties which characterize phase change materials. In particular, it will be shown that the crystalline state of phase change materials is characterized by the occurrence of resonant bonding, a particular favour of covalent bonding. This insight is employed to predict systematic property trends and to develop non-volatile memories with DRAM-like switching speeds potentially paving the road towards a universal memory. Phase change materials do not only provide exciting opportunities for applications including 'greener' storage devices, but also form a unique quantum state of matter as will be demonstrated by transport measurements. In this talk, potential limits of phase change memories in terms of switching speed, scalability and power consumption will be discussed.
Keywords - Phase change materials, non-volatile memory

[276] Magnetic Memories: From DRAM Replacement to Ultra Low Power Logic Chips
The recent advent of spin transfer torque (STT) has shed a new light on MRAM with the promises of much improved performances and greater scalability to very advanced technology nodes. As a result, MRAM is now viewed as a credible solution for stand-alone and embedded applications where the combination of non-volatility, speed and endurance is key. Whereas the technology is nearing maturity for DRAM replacement, with the exception of process scaling to sub-20nm which remains a challenge, circuit designers are now actively looking at SoCs where MRAM could bring in better performance and lower-power consumption in data intensive applications as well as instant-on capability in mobile applications. In this paper we present a review of the MRAM technology and a methodology for ASIC design using a custom full digital hybrid CMOS/Magnetic Process Design Kit. We finish by a few examples showing that magnetic memories can be efficiently integrated in logic designs, for both safety and low power purposes.
Keywords - spin transfer torque, MRAM, hybrid CMOS

[277] Resistive Memories: Which Applications?
Fabien Clermidy, Natalija Jovanovic, Santhosh Onkaraiha, Hocine Ouachekhi, Olivier Thomas, Ogun Turkyilmaz, Elisa Vianello, Jean-Michel Portal, Marc Bouquet
Recent announcement of 16Gbit Resistive memory from Sony shows the trend to quickly adopt resistive memories as an alternative to DRAM. However, using ReRAM for embedded computing is still a futuristic goal. This paper approaches two applications based on ReRAM-devices for gaining area, performance or power consumption. The first application is FPGA, one of the first architecture that can benefit the most from ReRAM integration to reduce footprint and save energy. The second application relates to ultra-low-power systems and the way to obtain an instantaneous "freeze" mode in devices for Internet of Things.

[278] Thinfilm Printed Ferro-Electric Memories and Integrated Products
Chistter Kaftson, Peter Fischer
Printed electronics has recently moved from a focus on the production of individual components towards the design and initial commercialization of integrated systems. This paper describes the current status and further trends of ferroelectric nonvolatile memories as developed and produced by Thin Film Electronics.
Keywords - Printed memory, ferroelectric, printed transistors, integrated products

10.2 Wireless NoCs
Chairs: Giorgos Dimitrakopoulos, Democontius University of Thrace, GR; Valenita Bertacco, University of Michigan, US

[279] An Adaptive Transmitting Power Technique for Energy Efficient mm-Wave Wireless NoCs
Andrea Mineo, Maurizio Palesi, Giuseppe Ascia, Vincenzo Catania
Several emerging techniques have been recently proposed for alleviating the communication latency and the energy consumption issues in multi/many-core architectures. One of such emerging communication techniques, namely, WiNoC replaces the traditional wired links with the use of wireless medium. Unfortunately, the energy consumed by the RF transceiver (i.e., the main building block of a WiNoC), and in particular by its transmitter, accounts for a significant fraction of the overall communication energy. In this paper we propose a runtime tunable transmitting power technique for improving the energy efficiency of the transceiver in wireless NoC architectures. The basic idea is tuning the transmitting power based on the location of the recipient of the current communication. The integration of this technique into two known WiNoC architectures, namely, WiNoC4 and MCWiNoC resulted in an energy reduction of 43% and 60%, respectively.

[280] Performance Evaluation of Wireless NoCs in Presence of Irregular Network Routing Strategies
Paul Wettin, Jacob Murray, Ryan Kim, Xinmin Yu, Pantha Pratim Panda, Deukhyun Heo
The millimeter (mm)-wave small-world wireless NoC (mSWNoC) is an enabling interconnect architecture to design high performance and low power multicore chips. As the mSWNoC has an overall irregular topology, it is extremely important to design suitable deadlock-free routing mechanisms for it. In this paper we quantify the latency, energy dissipation, and thermal profiles of mSWNoC architectures by incorporating irregular network routing strategies. We demonstrate that the latency, energy dissipation, and thermal profile are affected by the adopted routing methodologies. In presence of the benchmarks considered, the variation in latency and energy dissipation is small. However, the network hotspot temperature can vary considerably depending on the exact routing strategy and the characteristics of the benchmark.
Keywords - millimeter-wave wireless; network-on-chip; smallworld; irregular; routing algorithms

[281] Low Latency Wireless 3D NoCs via Randomized Shortcut Chips
Hiroki Matsuaji, Michihiro Kobuchi, Iki Fujiwara, Takahito Kagaami, Yasuhito Take, Tadahito Kuroda, Paul Bogdan, Rudar Marulescu, Hdehara Amano
In this paper, we demonstrate that we can reduce the communication latency significantly by inserting a fraction of randomness into a wireless 3D NoC (where CMOS wireless links are used for vertical inter-chip communication) when considering the physical constraints of the 3D design space. Towards this end, we consider two cases, namely 1) replacing existing horizontal 2D links in a wireless 3D NoC with randomized shortcut NoC links and 2) enabling full connectivity by adding a randomized NoC layer to a wireless 3D platform with partial or no horizontal
Connectivity. Consequently, the packet routing is optimized by exploiting both the existing and the newly added random NoC. At the same time, by adding randomly wired shortcut NoCs to a wireless 3D platform, a good balance can be established between the modularity of the design and the minimum randomness needed to achieve low latency, and experimental results show that by adding a random NoC chip to wireless 3D CMPS without built-in horizontal connectivity, the communication latency can be reduced by as much as 26.2% when compared to adding a 2D mesh NoC. Also, the application execution time and average fit transfer energy can be improved accordingly.

Global Fan Speed Control Considering Non-Ideal Temperature Measurements in Enterprise Servers
Jungsoo Kim, Mohamed M. Sabry, David Atenza, Kaiyan Vaidyanathan, Kenny Gross

Time lag and quantization in temperature sensors in enterprise servers lead to stability concerns on existing variable fan speed control schemes. Stability challenges become further aggravated when multiple local controllers are running together with the fan control scheme. In this paper, we present a global control scheme which tackles the concerns on the stability of enterprise servers while reducing the performance degradation caused by the variable fan speed control scheme. We first present a stable fan speed control scheme based on the Proportional-Integral-Derivative (PID) controller by adaptively adjusting the PID parameters according to the operating fan speed and eliminating the fan speed oscillation caused by temperature quantization. Then, we present a global control scheme which coordinates control actions among multiple local controllers. In addition, it guarantees the server stability while minimizing the overall performance degradation. We validated the proposed control scheme using a presently shipping commercial enterprise server. Our experimental results show that the proposed fan control scheme is stable under the non-ideal temperature measurement system (10 sec in time lag and 1°C in quantization figures). Furthermore, the global control scheme enables to run multiple local controllers in a stable manner while reducing the performance degradation up to 18.2% compared to conventional coordination schemes with 19.1% savings in power consumption.

Unveiling Eurora: Thermal and Power Characterization of the Most Energy-Efficient Supercomputer in the World
Andrea Bartolini, Matteo Cacciari, Carlo Cavazzoni, Giampietro Tecchiolli, Luca Benini

Eurora (EUropean many integrated CoR Architectures) is today the most energy efficient supercomputer in the world. Ranked 1st in the Green500 in July 2013, it is a prototype built from Eurotech and Cineca toward next-generation Tier-0 systems in the PRACE 2IP EU project. Eurora's outstanding energy-efficiency is achieved by adopting a direct liquid cooling solution and a heterogeneous architecture with best-in-class general purpose HW components (Intel Xeon E5, Intel Xeon Phi and NVIDIA Kepler K20). In this paper we present a novel, low-overhead monitoring infrastructure capable to track in detail and in real-time the thermal and power characteristics of Eurora's components with fine-grained resolution. Furthermore, the global control scheme enables to run multiple local controllers in a stable manner while reducing the performance degradation up to 18.2% compared to conventional coordination schemes with 19.1% savings in power consumption.

Contention Aware Frequency Scaling on CMPs with Guaranteed Quality of Service
Hao Shen, Qinru Qi

Workload consolidation is usually performed in datacenters to improve server utilization for higher energy efficiency. One of the key issues related to workload consolidation is contention for shared resources such as last level cache, main memory, memory controller, etc. Dynamic voltage and frequency scaling (DVFS) of CPU is another effective technique that has widely been used to trade the performance for power reduction. We have found that the degree of resource contention of a system affects its performance sensitivity to CPU frequency. In this paper, we apply machine learning techniques to construct a model that quantifies runtime performance degradation caused by resource contention and frequency scaling. The inputs of our model are readings from Performance Monitoring Units (PMUs) screened using standard feature selection technique. The model is tested on an SMP-enabled chip multiprocessor and it reaches up to 90% accuracy. Experimental results show that, guided by the performance model, runtime power management techniques such as DVFS can achieve more accurate power and performance tradeoff without violating the quality of service (QoS) agreement. The QoS violation of the proposed system is significantly lower than systems that have no performance degradation information.

Key words: contention, frequency scaling, power management, contention

Hybrid Wire-Surface Wave Architecture for One-to-Many Communication in Network-on-Chip
Ammar Kaskar, Noz Dahir, Ra’ed Al-Dujaily, Kenneth Tong, Ternence Mak, Alex Yakovlev

Network-on-chip (NoC) is a communication paradigm that has emerged to tackle different on-chip challenges and has satisfied different demands in terms of high performance and economical interconnect implementation. However, merely metal based NoC pursuit offered limited scalability with the relentless technology scaling, especially in one-to-many (1-to-M) communication. To meet the scalability demand, this paper proposes a new hybrid architecture empowered by both metal interconnects and Zenneck surface wave interconnects (SWI).

This architecture, in conjunction with newly proposed routing and global arbitration schemes, avoids overloading the NoC and alleviates traffic hotspots compared to the trend of handling 1-to-M traffic as unicast. This work addresses the system level challenges for intra chip multicasting. Evaluation results, based on a cycle-accurate simulation and hardware description, demonstrate the effectiveness of the proposed architecture in terms of power reduction ratio of 4 to 12x and average delay reduction of 25X or more, compared to a regular NoC. These results are achieved with negligible hardware overheads.

10.3 Green Computing Systems

Failure Analysis of a Network-on-Chip for Real-Time Mixed-Critical Systems
Eberle A Rambo, Alexander Tschelme, Jonas Dierer, Leonie Ahrendts, Rolf Ernst

Multi- and many-core architectures using Networks-on-Chip (NoC) are being explored for use in real-time safety-critical applications for their performance and efficiency. Such systems must provide isolation between tasks that may present distinct criticality levels. The NoC is critical to maintain the isolation property as it is a heavily used shared resource. To meet safety-standard requirements, such architectures require a systematic evaluation of the effects of all possible failures such as in the form of a Failure Mode and Effects Analysis (FMEA). We present the results of a detailed system-level analysis of a typical real-time mixed-critical network-on-chip architecture. This comprises an FMEA and error effects classification regarding duration and isolation violation.

Hybrid Wire-Surface Wave Architecture for One-to-Many Communication in Network-on-Chip
Ammar Kaskar, Noz Dahir, Ra’ed Al-Dujaily, Kenneth Tong, Ternence Mak, Alex Yakovlev

Network-on-chip (NoC) is a communication paradigm that has emerged to tackle different on-chip challenges and has satisfied different demands in terms of high performance and economical interconnect implementation. However, merely metal based NoC pursuit offered limited scalability with the relentless technology scaling, especially in one-to-many (1-to-M) communication. To meet the scalability demand, this paper proposes a new hybrid architecture empowered by both metal interconnects and Zenneck surface wave interconnects (SWI).

This architecture, in conjunction with newly proposed routing and global arbitration schemes, avoids overloading the NoC and alleviates traffic hotspots compared to the trend of handling 1-to-M traffic as unicast. This work addresses the system level challenges for intra chip multicasting. Evaluation results, based on a cycle-accurate simulation and hardware description, demonstrate the effectiveness of the proposed architecture in terms of power reduction ratio of 4 to 12x and average delay reduction of 25X or more, compared to a regular NoC. These results are achieved with negligible hardware overheads.

Unveiling Eurora: Thermal and Power Characterization of the Most Energy-Efficient Supercomputer in the World
Andrea Bartolini, Matteo Cacciari, Carlo Cavazzoni, Giampietro Tecchiolli, Luca Benini

Eurora (EUropean many integrated CoR Architectures) is today the most energy efficient supercomputer in the world. Ranked 1st in the Green500 in July 2013, it is a prototype built from Eurotech and Cineca toward next-generation Tier-0 systems in the PRACE 2IP EU project. Eurora's outstanding energy-efficiency is achieved by adopting a direct liquid cooling solution and a heterogeneous architecture with best-in-class general purpose HW components (Intel Xeon E5, Intel Xeon Phi and NVIDIA Kepler K20). In this paper we present a novel, low-overhead monitoring infrastructure capable to track in detail and in real-time the thermal and power characteristics of Eurora's components with fine-grained resolution. Furthermore, the global control scheme enables to run multiple local controllers in a stable manner while reducing the performance degradation up to 18.2% compared to conventional coordination schemes with 19.1% savings in power consumption.

Unveiling Eurora: Thermal and Power Characterization of the Most Energy-Efficient Supercomputer in the World
Andrea Bartolini, Matteo Cacciari, Carlo Cavazzoni, Giampietro Tecchiolli, Luca Benini

Eurora (EUropean many integrated CoR Architectures) is today the most energy efficient supercomputer in the world. Ranked 1st in the Green500 in July 2013, it is a prototype built from Eurotech and Cineca toward next-generation Tier-0 systems in the PRACE 2IP EU project. Eurora's outstanding energy-efficiency is achieved by adopting a direct liquid cooling solution and a heterogeneous architecture with best-in-class general purpose HW components (Intel Xeon E5, Intel Xeon Phi and NVIDIA Kepler K20). In this paper we present a novel, low-overhead monitoring infrastructure capable to track in detail and in real-time the thermal and power characteristics of Eurora's components with fine-grained resolution. Furthermore, the global control scheme enables to run multiple local controllers in a stable manner while reducing the performance degradation up to 18.2% compared to conventional coordination schemes with 19.1% savings in power consumption.

Contention Aware Frequency Scaling on CMPs with Guaranteed Quality of Service
Hao Shen, Qinru Qi

Workload consolidation is usually performed in datacenters to improve server utilization for higher energy efficiency. One of the key issues related to workload consolidation is contention for shared resources such as last level cache, main memory, memory controller, etc. Dynamic voltage and frequency scaling (DVFS) of CPU is another effective technique that has widely been used to trade the trade the performance for power reduction. We have found that the degree of resource contention of a system affects its performance sensitivity to CPU frequency. In this paper, we apply machine learning techniques to construct a model that quantifies runtime performance degradation caused by resource contention and frequency scaling. The inputs of our model are readings from Performance Monitoring Units (PMUs) screened using standard feature selection technique. The model is tested on an SMP-enabled chip multiprocessor and it reaches up to 90% accuracy. Experimental results show that, guided by the performance model, runtime power management techniques such as DVFS can achieve more accurate power and performance tradeoff without violating the quality of service (QoS) agreement. The QoS violation of the proposed system is significantly lower than systems that have no performance degradation information.

Key words: contention, frequency scaling, power management, contention

COOLIP - Simple yet Effective Job Allocation for Distributed Thermally-Throttled Processors
Pratyush Kumar, Hoesook Yang, Iuliana Bacivarov, Lothar Thiele

Thermal constraints limit the time for which a processor can run at high frequency. Such thermal-throttling complicates the computation of response times of jobs. For multiple processors, a key decision is where to allocate the next job. For distributed thermally-throttled processors, we present COOLIP with a simple allocation policy: a job is allocated to the earliest available processor, and if there are several available simultaneously, to the coolest one. For Poisson distribution of inter-arrival times and Gaussian distribution of execution demand of jobs, COOLIP matches the 95-percentile response time of Earliest Finish-Time (EFT) policy which minimizes response time with full knowledge of execution demand of unfinished jobs and
thermal models of processors. We argue that COOLIP performs well because it directs the processors into states such that a defined sufficient condition of optimality holds.

Efficient Simulation and Modelling of Non-rectangular NoC Topologies
Ji Qi, Mark Zwolinski
With increasing chip complexity, Networks-on-Chips (NoCs) are becoming a central platform for future on-chip communications. Many regular NoC architectures have been proposed to eliminate the communication bottlenecks of traditional bus-based networks. Non-rectangular and irregular architectures have also been proposed to increase performance. However, the complexity of designing custom non-rectangular networks leads to a rapid increase in design and verification times. To alleviate the conflict between performance and efficiency, this paper proposes a novel method that efficiently constructs virtual non-rectangular topologies on a mesh network by using time-regulated models to emulate irregular patterns. Data routings on virtual hexagonal and two irregular geometries validate the proposed method. An MPEG-4 decoder is used to exemplify the proposed method for media applications. Results analysis shows the virtual topologies emulated by the proposed method can provide precise timing and energy performance.

Moving from Co-Simulation to Simulation for Effective Smart Systems Design
Franco Fummi, Michele Lora, Francesco Stefanni, Dimitrios Tsachani, Jan Vanhese, Sara Vinco
Design of smart systems needs to cover a wide variety of domains, ranging from analogue to digital, with power devices, micro-sensors and actuators, up to MEMS. This high level of heterogeneity makes design a very challenging task, as each domain is supported by specific languages, modeling formalisms and simulation frameworks. A major issue is furthermore posed by simulation, that heavily impacts the design and verification loop and that is very hard to be built in such an heterogeneous context. On the other hand, achieving efficient simulation would indeed make system design feasible with respect to budget constraints. This work provides a formalization of the typical abstraction levels and design domains of a smart system. This taxonomy allows to identify a precise role in the design flow for co-simulation and simulation scenarios. Moreover, a methodology is proposed to move from the co-simulated heterogeneous to a simulatable homogeneous representation in C++ of the entire smart system. The impact of heterogeneous or homogeneous models of computation is also examined. Experimental results prove the effectiveness of the proposed C++ generation for reaching high-speed simulation.

Hardware-Based Fast Exploration of Cache Hierarchies in Application Specific MPSoCs
Itsko Navinne, Josef Schneider, Hars Javadi, Sri Parameswaran
Multi-level caches are widely used to improve the memory access speed of multiprocessor systems. Deciding on a suitable set of cache memories for an application specific embedded system's memory hierarchy is a tedious problem, particularly in the case of MPSoCs. To accurately determine the number of hits and misses for all the configurations in the design space of an MPSoC, researchers extract the trace first using Instruction set simulators and then simulate using a software simulator. Such simulations take several hours to months. We propose a novel method based on specialized hardware which can quickly simulate the design space of cache configurations for a shared memory multiprocessor system on an FPGA, by analyzing the memory traces and calculating the cache hits and miss simultaneously. We demonstrate that our simulator can explore the cache design space of a quad-core system with private L1 caches and a shared L2 cache, over a range of standard benchmarks, taking as less as 0.106 seconds per million memory accesses, which is up to 456 times faster than the fastest known software based simulator. Since we emulate the program and analyze memory traces simultaneously, we eliminate the need to extract multiple memory access traces prior to simulation, which saves a significant amount of time during the design stage.

Energy Optimization in 3D MPSoCs with Wide-I/O DRAM Using Temperature Variation Aware Bank-wise Refresh
Mohammadreza Sadri; Matthias Jung; Christian Wee; Norbert Wehn; Luca Benni
Heterogeneous 3D integrated systems with Wide-I/O DRAMs are a promising solution to squeeze more functionality and storage bits into an ever decreasing volume. Unfortunately, with 3D stacking, the challenges of high power densities and thermal dissipation are exacerbated. We improve DRAM refresh power by considering the lateral and vertical temperature variations in the 3D structure and adapting the per-DRAM-bank refresh period accordingly. In order to provide proof of our concepts we develop an advanced virtual platform which models the performance, power, and thermal behavior of a 3D-integrated MPSoC with Wide-I/O DRAMs in detail. On this platform we run the Android OS with real-world benchmarks to quantify the advantages of our ideas. We show improvements of 16% in DRAM refresh power due to temperature variation aware bank-wise refresh. Furthermore, two solutions are investigated to speedup system simulations: (1) Adaptive tuning of sampling intervals based on the estimated chip thermal profile, which results in speedups of 2X. (2) Hardware acceleration of thermal simulations using the Mixeler engine, which shows possible speedups of 12X.

10. System-level Evaluation

Chairs: Pablo Sanchez, University of Cantabria, ES; Florian Lethome, Synopsys, FR

Automatic Detection of Concurrency Bugs through Event Ordering Constraints
Luis Gabriel Munilo, Simon Wawrowsches, Jeronimo Castilld, Rainer Leupers, Gerdi Ascheid
Writing correct parallel software for modern multiprocessor systems-on-chip (MPSoCs) is a complicated task. Programmers can rarely anticipate all possible external and internal interactions in complex concurrent systems. Concurrency bugs originating from races and improper synchronization are difficult to understand and reproduce. Furthermore, traditional debug and verification practices for embedded systems lack support to address this issue efficiently. For instance, programmers still need to step through several executions until finding a buggy state or analyze complex traces, which results in productivity losses. This paper proposes a new debug approach for MPSoCs that combines dynamic analysis and the benefits of virtual platforms. All in all, it (i) enables automatic exploration of SW behavior, (ii) identifies problematic concurrent interactions, (iii) provokes possibly erroneous executions and, ultimately, (iv) detects concurrency bugs. The approach is demonstrated on an industrial strength virtual platform with a full Linux operating system and real-world parallel benchmarks.

Lorenzo Zunilo, Cristian Zambelli, Rino Micheloni, Salvatore Galliano, Marco Indaco, Stefano Di Carlo, Paolo Prinetto, Pietro Olivo, Davide Bertocci
Solid State Drives (SSDs) are gaining particular momentum in various frameworks such as multimedia, large data centers and cloud environments. Unfortunately, efficient CAD tools for SSD design space exploration able to assess the optimization of the device microarchitecture w.r.t. the target performance are still missing. This paper tries to close this gap by proposing SSDExplorer, a tool for fine-grained and fast design space exploration of SSD devices. SSDExplorer provides unprecedented insights into the architecture behavior and subcomponent interaction efficiency, while avoiding the need for the actual implementation of an FTL or of key hardware components. This is achieved by the introduction of suitable abstractions of the different components. This is confirmed by the thorough validation of SSDExplorer against a commercial SSD device.

May-Happen-in-Parallel Analysis Based on Segment Graphs for Safe ESL Models
Wenwei Chen, Xu Han, Rainer Dömer
A well-defined system-level model contains explicit parallelism and should be free from parallel access conflicts to shared variables. However, safe parallelism is difficult to achieve since risky shared variables are often hidden in the design and are not exposed through simulation. In this paper, we propose a new static analysis approach based on segment graphs that identifies a tight set of potential access conflicts in segments that may-happen-in-parallel (MHP). Our experimental results show that the analysis is complete, accurate and fast to reveal dangerous shared variables in several embedded application models. Compared to earlier works, our approach significantly reduces the number of false conflict reports and thus saves the designer time.

10.5 Analysis of Components and Systems

Chairs: Frank Oppenheimer, OFFIS, DE; Todor Stefanov, Leiden University, NL

Efficient Simulation and Modelling of Non-rectangular NoC Topologies
Ji Qi, Mark Zwolinski
With increasing chip complexity, Networks-on-Chips (NoCs) are becoming a central platform for future on-chip communications. Many regular NoC architectures have been proposed to eliminate the communication bottlenecks of traditional bus-based networks. Non-rectangular and irregular architectures have also been proposed to increase performance. However, the complexity of designing custom non-rectangular networks leads to a rapid increase in design and verification times. To alleviate the conflict between performance and efficiency, this paper proposes a novel method that efficiently constructs virtual non-rectangular topologies on a mesh network by using time-regulated models to emulate irregular patterns. Data routings on virtual hexagonal and two irregular geometries validate the proposed method. An MPEG-4 decoder is used to exemplify the proposed method for media applications. Results analysis shows the virtual topologies emulated by the proposed method can provide precise timing and energy performance.

Moving from Co-Simulation to Simulation for Effective Smart Systems Design
Franco Fummi, Michele Lora, Francesco Stefanni, Dimitrios Tsachani, Jan Vanhese, Sara Vinco
Design of smart systems needs to cover a wide variety of domains, ranging from analogue to digital, with power devices, micro-sensors and actuators, up to MEMS. This high level of heterogeneity makes design a very challenging task, as each domain is supported by specific languages, modeling formalisms and simulation frameworks. A major issue is furthermore posed by simulation, that heavily impacts the design and verification loop and that is very hard to be built in such an heterogeneous context. On the other hand, achieving efficient simulation would indeed make system design feasible with respect to budget constraints. This work provides a formalization of the typical abstraction levels and design domains of a smart system. This taxonomy allows to identify a precise role in the design flow for co-simulation and simulation scenarios. Moreover, a methodology is proposed to move from the co-simulated heterogeneous to a simulatable homogeneous representation in C++ of the entire smart system. The impact of heterogeneous or homogeneous models of computation is also examined. Experimental results prove the effectiveness of the proposed C++ generation for reaching high-speed simulation.
10.6 Multi-processor and Distributed Systems

Chair: Orlando Moreira, Ericsson, NL; Benny Akeson, CTU, CZ

[298] Thermal-Aware Frequency Scaling for Adaptive Workloads on Heterogeneous MPSoCs
Heng Yu, Rizwan Syed, Yajun Ha

For applications featuring adaptive workloads, the quality of their task execution can be dynamically adjusted given the runtime constraints. When mapping them to heterogeneous MPSoC, it is expected not only to achieve the highest possible execution quality, but also meet the critical thermal challenges from the continuously increasing chip density. Prior thermal management techniques, such as Dynamic Voltage/Frequency Scaling (DVFS) and thread migration, do not take into account the trade-off possibility between execution quality and temperature control. In this paper, we explore the capability of adaptive workloads for effective temperature control, while maximally ensuring the execution Quality-of-Service (QoS). We present a thermal-aware dynamic frequency scaling (DFS) algorithm on heterogeneous MPSoCs, where judicious frequency selection achieves QoS maximization under the temperature threshold, which is converted to the thermal timing deadline as an additional execution constraint. Results show that our frequency scaling algorithm achieves as large as 31.5% execution cycle/QoS improvement under thermal constraints.

[300] Partitioned Mixed-Critically Scheduling on Multiprocessor Platforms
Chuancai Gu, Nan Quinn, Qingru Deng, Wang Yi

Scheduling mixed-critically systems that integrate multiple functionalities with different criticality levels into a shared platform appears to be a challenging problem, even on single-processor platforms. Multicore processors are more and more widely used in embedded systems, which provide greater opportunities for such mixed-critically systems. In this paper, we propose a partitioned scheduling algorithm MPVD to extend the state-of-the-art single-processor mixed-critically scheduling algorithm EY to multiprocessor platforms. The key idea of MPVD is to evenly allocate tasks with different criticality levels to different processors, in order to better explore the asymmetry between different criticality levels and improve the system schedulability. Then we propose two enhancements to further improve the schedulability of MPVD. Experiments with randomly generated task sets show significant performance improvement of our proposed approach over existing algorithms.

[301] Generation of Communication Schedules for Multi-Mode Distributed Real-Time Applications
Akrumal Azim, Gonzalo Carvajal, Rodolfo Pellizzoni, Sebastian Fischmeister

A key problem in designing multi-mode real-time systems is the generation of schedules to reduce the complexities of transforming the model semantics to code. Moreover, distributed multi-mode applications are prone to suffer from delays incurred during mode changes. We therefore aim to generate communication schedules that have low average mode-change delay for multi-mode real-time distributed applications. In this paper, we use optimization constraints associated to timing requirements to generate state-based schedules for multi-mode communication systems, and illustrate the workflow for generating schedules from specifications through a real-time video monitoring case-study. Our experiments in the case-study demonstrate that schedules generated using the proposed method reduce the average mode-change delay in relation to a randomized algorithm and the well-known EDF scheduling algorithm.

10.7 Advances in Synthesis

Chair: John Hayes, University of Michigan, US; Kim Taemin, Intel Labs, US

[302] Provably Minimal Energy Using Coordinated DVFS and Power Gating
Nathaniel A. Conos, Sara Meguerdichian, FOad Gabr, Midrag Potkonjak

Both energy and execution speed can be greatly impacted by clock and power gating, nonlinear voltage scaling, and leakage energy. We address the problem of coordinated power gating and dynamic voltage scaling (DVS) to minimize the overall energy consumption of an application under user-specified timing constraints. We prove that a solution provided by our convex programming formulation that uses at most two versions of hardware, where each version uses its own constant voltages, is optimal. Comprehensive evaluation of the new approach demonstrates energy improvements over traditional DVS and DVS and power gating techniques by factors of 1.44X-2.97X and 1.44X-2.82X, respectively.

[303] A Tree Arbiter Cell for High-Speed Resource Sharing in Asynchronous Environments
Syed Rameez Naqvi, Andreas Steininger

We present a novel tree arbiter cell that allows a pipelined processing of asynchronous requests. In this way it can achieve significantly lower delay in the critical case of frequent requests.
coming from different clients. We elaborate the necessary extension to facilitate a cascaded use of this cell in a tree-like fashion, and we show by theoretical analysis that in this configuration our cell provides better fairness than the standard approach. We implement our approach and quantitatively compare its performance properties with related work in a gate-level simulation. In our sample asynchronous Networks-on-Chip application our new cell proves to increase the throughput of three different designs available in literature by approximately 61.28%, 69.24%, and 186.85% respectively.

[304] An Efficient Manipulation Package for Biconditional Binary Decision Diagrams
Luca Amari, Pierre-Emmanuel Gallardon, Giovanni De Micheli

Biconditional Binary Decision Diagrams (BBDDs) are a novel class of binary decision diagrams where the branching condition, and its associated logic expansion, is biconditional on two variables. Reduced and ordered BBDDs are remarkably compact and unique for a given Boolean function. In order to exploit BBDDs in Electronic Design Automation (EDA) applications, efficient manipulation algorithms must be developed and integrated in a software package. In this paper, we present the theory for efficient BBDD manipulation and its practical software implementation. The key features of the proposed approach are (i) strong canonical form pre-conditioning of stored BBDD nodes, (ii) recursive formulation of Boolean operations in terms of biconditional expansions, (iii) performance-oriented memory management and (iv) dedicated BBDD re-ordering techniques. Experimental results show that the developed BBDD package achieves an average node count reduction of 19.48% and a speed-up factor of 1.63x with respect to a state-of-art decision diagram manipulation package. Employed in the synthesis of datapath circuits, the BBDD manipulation package is capable to advantageously restructure arithmetic operations producing 11.02% smaller and 32.29% faster circuits as compared to a commercial synthesis flow.

[305] Synthesis Algorithm of Parallel Index Generation Units
Yusuke Matsunaga

The index generation function is a multi-valued logic function which checks if the given input vector is a registered or not, and returns its index value if the vector is registered. If the latency of the operation is critical, dedicated hardware is used for implementing the index generation functions. This paper proposes a method implementing the index generation functions using parallel index generation units. A novel and efficient algorithm called ‘conflict free partitioning’ is proposed to synthesis parallel index generation units. Experimental results show the proposed method outperforms existing methods.

Keywords - Index generation function, logic synthesis

[306] Automating Data Reuse in High-Level Synthesis
Wim Meeus, Dirk Strickland

Current High-Level Synthesis (HLS) tools perform excellently for the synthesis of computation kernels, but they often don't optimize memory bandwidth. As memory access is a bottleneck in many algorithms, the performance of the generated circuit will benefit substantially from memory access optimization. In this paper we present an automated method and a toolchain to detect reuse of array data in loop nests and to build hardware that exploits this data reuse. This saves memory bandwidth and improves circuit performance. We make use of the polyhedral representation of the source program, which makes our method computationally easy. Our software complements the existing HLS flows. Starting from a loop nest written in C, our tool generates a reuse buffer and a loop controller, and preprocesses the loop body for synthesis with an existing HLS tool. Our automated tool produces designs from unoptimized source code that are as efficient as those generated by a commercial HLS tool from manually-optimized source code.

[307] A Universal Symmetry Detection Algorithm
Peter M. Maurer

Research on symmetry detection focuses on identifying and detecting new types of symmetry. We present an algorithm that is capable of detecting any type of permutation-based symmetry, including many types for which there are no existing algorithms. General symmetry detection is library-based, but symmetries that can be parameterized, (i.e. total, partial, rotational, and dihedral symmetry), can be detected without using libraries. In many cases it is faster than existing techniques. Furthermore, it is simpler than most existing techniques, and can easily be incorporated into existing software.

[308] Optimization of Design Complexity in Time-Multiplexed Constant Multiplications
Levent Aksoy, Paulo Flores, José Monteiro

The multiplication of constants by a data input is an essential operation in digital signal processing (DSP) systems. For applications requiring a large number of constant multiplications under stringent hardware constraints, it is generally realized under a folded architecture, where a single constant selected from a set of multiple constants is multiplied by the data input at each time, called time-multiplexed constant multiplication (TMCM). This paper addresses the problem of optimizing the complexity of a TMCM design and introduces an algorithm that finds the least complex TMCM design by sharing the logic operators, i.e., adders, subtractors, adders/subtractors, and multiplexors (MUXes). It includes efficient search methods, yielding better results than existing TMCM algorithms.

[309] Hardware Primitives for the Synthesis of Multithreaded Elastic Systems

Elastic systems operate in a dataflow-like mode using a distributed scalable control and tolerating variable-latency computations. At the same time, multithreading increases the utilization of processing units and hides the latency of each operation by time-multiplexing operations of different threads in the datapath. This paper proposes a model to unify multithreading and elasticity. A new multithreaded elastic control protocol is introduced supported by low-cost elastic buffers that minimize the storage requirements without sacrificing performance. To enable the synthesis of multithreaded elastic architectures, new hardware primitives are proposed and utilized in two circuit examples to prove the applicability of the proposed approach.

11.0 Special Day Keynote: Organic Electronics - From Lab to Markets

Organizers: Ian O'Connor, Lyon Institute of Nanotechnology, FR; Thomas Mikolajick, NamLab gGmbH, DE
Chairs: Ian O'Connor, Lyon Institute of Nanotechnology, FR; Thomas Mikolajick, NamLab gGmbH, DE

11.1 Embedded Tutorial: Alternatives to CMOS

Organizers: Ian O'Connor, Lyon Institute of Nanotechnology, FR; Thomas Mikolajick, NamLab gGmbH, DE
Chairs: Ian O'Connor, Lyon Institute of Nanotechnology, FR; Thomas Mikolajick, NamLab gGmbH, DE

[310] Spintronics for Low-Power Computing
Yue Zhang, Weisheng Zhao, Jacques-Olivier Klein, Wang Kang, Damien Querloz, Youguang Zhang, Dafne Ravelosona, Claude Chappert

Microelectronics has been following Moore's law for almost 49 years. However this trend tends to run out of steam in recent technology nodes. The continuous improvements in the size of the transistors and in the operating frequencies result in serious power consumption, heat dissipation and reliability issues. Spintronics (Nobel Prize of Physics, 2007 awarded to Prof. Fert from Univ. Paris-Sud and Peter Grünberg from Forschungszentrum Jülich) nanodevices can reduce significantly the power, improve the reliability or allow new functionalities. The 2010 ITRS report on emerging research devices identified Magnetic Tunnel Junction (MTJ) nanopillar (the prevalent spintronics nanodevice) as one of the most promising technologies to be part of the future mesocircuits. It provides data non-volatility, hardiness to radiations, fast data access and low-power operations. Magnetic memories become the most promising
candidate for both low power logic computing and the data storage. This tutorial paper presents multi-discipline questions (Device, Circuit, Architecture, and System and CAD) related to this topic to share the most recent results and discuss the future challenges.

Sébastien Le Beux, Hui Li, Ian O'Connor, Kazem Chehmi, Xuchen Liu, Jelena Trajkovic, Gabriela Nicolescu
The next generation of MPSoC points to the integration of thousands of IP cores, requiring high performance interconnect for high throughput communications. Optical on-chip interconnect enables significantly increased bandwidth and decreased latency in MPSoC. However, the interface between electrical and photonic devices implies strong layout constraints that may impact the system performance and scalability. In this paper, we propose a novel optical interconnect named CHAMELEON. The interface simplifies the layout and allows the bandwidth between IP cores to be adapted according to the communication requirements. Compared to related networks, CHAMELEON demonstrates improved scalability and flexibility at the cost of minor increase in power consumption.

Keywords - Optical Network on Chip; MPSoC, WDM.

Ute Zschieschang, Reinhold Ablod, Ulrike Kraft, Kazuo Takimya, Tarek Zak, Florian Letzkus, Jörg Butschke, Harald Richter, Joachim N. Burghartz, Wei Xiong, Boris Mummann, Hagen Klauk
A process for the fabrication of bottom-gate, top-contact (inverted staggered) organic thin-film transistors (TFTs) with channel lengths as short as 1 μm on flexible plastic substrates has been developed. The TFTs employ vacuum-deposited small molecule semiconductors and a low-temperature processed gate dielectric that is sufficiently thin to allow the TFTs to operate with voltages of about 3 V. The p-channel TFTs have an effective field-effect mobility of about 1 cm²/Vs, an on/off ratio of 10⁷, and a signal propagation delay (measured in 11-stage ring oscillators) of 300 ns per stage. For the n-channel TFTs, an effective field-effect mobility of about 0.06 cm²/Vs, an on/off ratio of 106, and a signal propagation delay of 17 μs per stage have been obtained.

11.2 Transitioning NoC Design Techniques to Future Challenges
Chairs: Masaoud Daneshtalab, University of Turku, FI; Heikki Matsutani, Keio University, JP

Dowell Lee, Ritesh Parkh, Valeria Bertacco
The expected low reliability of the silicon substrate at upcoming technology nodes presents a key challenge for digital system designers. Networks-on-chip (NoCs) are especially concerning because they are often the only communication infrastructure for the chips in which they are deployed. Recently, routing reconfiguration solutions have been proposed to address this problem. However, they come at a high silicon cost, and often require suspending the normal network activity while executing a centralized, resource-hungry reconfiguration algorithm. This paper proposes a novel, fast and minimalistic routing reconfiguration algorithm, called BLINC. BLINC utilizes precomputed routing metadata to quickly evaluate localized detours upon each fault manifestation. We showcase the efficacy of our algorithm by deploying it in a novel NoC fault detection and reconfiguration solution, where BLINC enables uninterrupted NoC operation during aggressive online testing. If a fault seems likely to occur, we circumvent it in advance with the aid of our BLINC reconfiguration algorithm. Experimental results show an 80% reduction in the average number of routers affected by a reconfiguration event, compared to state-of-the-art techniques. BLINC enables negligible performance degradation in our detection and reconfiguration solution, while solutions based on current techniques suffer a 17-fold latency increase.

Tiansheng Zhang, Jose L. Abelein, Ajay Joshi, Ayse K. Coskun
Silicon-photonics on-chip (NoCs) provide high bandwidth density; therefore, they are promising candidates to replace electrical NoCs in manycore systems. The silicon-photonics NoCs, however, are sensitive to the temperature gradients that typically occur on the chip, and hence, require proactive thermal management. This paper first provides a design space exploration of silicon-photonics networks in manycore systems and quantifies the performance impact of the temperature gradients for various network bandwidths. The paper then introduces a novel job allocation technique that minimizes the temperature gradients among the ring modulators/filters to improve the application performance. Experimental results for a single-chip 256-core system demonstrate that our policy is able to maintain the maximum network bandwidth. Compared to existing workload allocation policies, the proposed policy improves system performance by up to 26.1% when running a single application and 18.3% for multi-program scenarios.

Luca Ramin, Paolo Gianni, Hervé Tatienguem Fankem, Alberto Ghinibaldi, Sandro Bartolini, Davide Bertozzi
Many crossbenchmarking results reported in the open literature raise optimistic expectations on the use of optical networks-on-chip (ONoCs) for high-performance and low-power on-chip communication. However, most of those previous works ultimately fail to make a compelling case for chip-level nanophotonic NoCs, especially for the lack of aggressive electronic baselines (EnOCs), and the poor accuracy in physical- and architecture-layer analysis of the ONoC. This paper aims at providing the guidelines and minimum requirements so that nanophotonic emerging technology may become of practical relevance. The key differentiating factor of this work consists of contrasting ONoC solutions with an aggressive ENoC architecture with realistic complexity, performance, and power figures, synthesized on an industrial 40nm low-power technology. At the same time, key physical design issues and network interface architecture requirements for the ONoC under test are carefully assessed, thus paving the way for a well-grounded definition of the requirements for the emerging ONoC technology to achieve the energy break-even point with respect to pure electronic interconnected solutions in future multi- and many-core systems.

Wolfgang Böter, Christof Oesewald, Daniel Gregorex, Alberto Garcia-Ortiz
The increasing requirements for bandwidth and quality-of-service motivate the use of parallel interconnect architectures with several degrees of reconfiguration. This paper presents an IP, called Distributed Channel Management (DCM), to extend existing packet-switched NoCs with a reconfigurable point-to-point network seamlessly, i.e., without the need for any modification on the routers. The configuration of the reconfigurable network takes place dynamically and autonomously, so that the topology can be changed at run time. Furthermore, the architecture is scalable due to the autonomous decentralized administration of the links. The Paper reports a thorough experimental analysis of the overhead of the approach at the gate level that considers different network parameters such as fit size and timing constraints.

Keywords - NoC, reconfigurable NoC, optical NoC, SoC

Gnaneswar Rao Joanna, John Jose, Rashana Radhakrishnan, Madhu Mutyam
With the drift from computation centric designs to communication centric designs in the Chip Multi Processor (CMP) era, the interconnect fabric is gaining more importance. An efficient NoC in terms of power, area and average flit latency has a huge impact on the overall performance of a CMP. In the current work, we propose MinBSD - a minimally buffered, single cycle, deflection router. It incorporates different operations (Injection, Ejection, Preemption, Re-injection) in a single module to handle the traffic effectively and ensures smooth flow of fits through router pipeline. It performs overlapped execution of independent operations. These factors not only make MinBSD to operate in a single cycle but also to reduce the critical path latency resulting in a faster interconnect network. Experimental results show that MinBSD reduces the average flit latency on real work loads, reduces die area and power consumption when compared to the existing state-of-the-art minimally buffered deflection routers.

11.3 Industry Relevant Research and Practice for System Design
Chairs: Emil Matus, Technische Universität Dresden, DE; Norbert Wehn, TU Kaiserslautern, DE

...
This paper presents an industry proven Metamodeling based approach to System-Level-Synthesis which is seen as generic design automation strategy above today’s implementation levels RTL (for digital) and Schematic Entry (for analog). The approach follows a new synthesis paradigm: The designer develops a simple domain and/or design specific language and a smart tool synthesizes implementation level models according to its needs. The overhead of making both a tool and a model pays off since the tool building is automated by code generation and reuse, both based on Metamodeling techniques. Also the focus on own demands keeps development costs low. Finally, specification data is utilized. i.e. the domain specific language simplifies to a document structure as a table. This keeps also modeling effort low since specification content is used and no model need to be built. Furthermore, increases design consistency and thus decreases debug time. Using these concepts, single design steps have been speed up to a factor of 20x and implementations of chips (specifications to tapeout) have been speed up to a factor of 3x.

Keywords - system level synthesis; metamodeling; code generation

For low-power digital ICs with ultra-wide voltage and frequency scaling (e.g., from the nominal supply voltage to the sub/near-threshold regime), achieving design closure can be a big challenge, especially when speed limits are pushed at very different voltages. This paper shares a practical logic synthesis recipe that helps to fulfill tight timing constraints. Our method includes: i) synthesizing circuits at high voltage; ii) over-constraining maximal transition time; iii) pruning standard cell library based on cell delay degradation factor across voltages. This approach shows effectiveness on an industrial 90nm low-power micro-controller.

Keywords - logic synthesis; ultra-wide voltage and frequency scaling; ultra-low power

To offer more computing power to modern SoCs, transistors keep scaling in new technology nodes. Consequently, the power density is increasing, leading to higher thermal risks. Thermal issues need to be addressed as early as possible in the design flow, when the optimization opportunities are the highest. For early design stages, architects rely on virtual prototypes to model their designs’ behavior with an adapted trade-off between accuracy and simulation speed. Unfortunately, accurate virtual prototypes fail to encompass thermal effects timescale. In this paper, we demonstrate that less accurate high-level architectural models, in conjunction with efficient power and thermal simulation tools, provide an adapted environment to analyze thermal issues and design software thermal mitigation solutions in the case of the Locomotive MPSoC architecture.

Keywords - system-level simulation; exploration of complex architectures; SoC

This paper presents our multi-year experience in the development of a Functional Modeling Compiler (FMC), a new model-based design tool for the development of multi-disciplinary automotive cyber-physical systems. We show how system-level simulation models suitable for design space exploration of complex architectures can be synthesized from functional specifications to test and validate the interactions between ECUs, control algorithms, and the multi-physics.

Keywords - parallel event-driven simulation; simulation; verification; SOC; EDA

Simulation-based techniques play a key role in validating the functional correctness of microprocessor designs. A common approach for validating microprocessors (called instruction-by-instruction, or IBI checking) consists of running an RTL and an architectural simulation in lock-step, while comparing processor architectural state at each instruction retirement. This solution, however, cannot be deployed on long regression tests, because of the limited performance of RTL simulators. Acceleration platforms have the performance power to overcome this issue, but are not amenable to the deployment of an IBI checking methodology. Indeed, validation on these platforms requires logging activity on-platform and then checking it against a golden model off-platform. Unfortunately, an IBI checking approach following this paradigm entails a large slowdown for the acceleration platform, because of the sizable amount of data that must be transferred off-platform for comparison against the golden model. In this work we propose a sequence-by-sequence (SBS) checking approach that is efficient and practical for acceleration platforms. Our solution validates the test execution over sequences of instructions (instead of individual ones), thus greatly reducing the amount of data transferred for off-platform checking. We found that SBS checking delivers the same bug-detection accuracy as traditional IBI checking, while reducing the amount of traced data by more than 90%.

Keywords - parallel event-driven simulation; simulation; verification; SoC; EDA
In post-silicon functional validation, one of the most complex and time-consuming processes is the localization of an instruction that exposes a bug detected at system level. The task is particularly difficult due to the silicon's limited observability and the long time between a failure's occurrence and its detection. We propose a novel method that automates the architectural localization of post-silicon test-case failures. Our proposed tool analyses a failing test-case, while leveraging the information derived from executing the test on an Instruction Set simulator (ISS), to identify a set of instructions that could lead to the faulty final state. The proposed failure localization process comprises the creation of a resource dependency graph based on the execution of the test-case on the ISS, determining a program slice of instructions that influence the faulty resources, and the reduction of the set of suspicious instructions by leveraging the knowledge of the correct resources. We evaluate our proposed solution through extensive experiments. Experimental results show that, in over 97% of all cases, our method was able to narrow down the list of suspicious instructions to under 2 instructions, on average, out of 200. In over 59% of all cases, our method correctly reduced a test-case to a single faulty instruction.

To observe internal signals, physical probing is an important step in post-silicon debug. Focused ion beam (FIB) is one of most popular probing technologies. However, an unsuitable layout significantly decreases the percentage of nets which can be observed through FIB probing for advanced process technologies. This paper presents the first design-for-debug routing to increase the FIB observable rate. The proposed algorithm, which adopts three FIB states and costs to enhance the maze routing, keeps at least one FIB candidate for each net while routing. Experimental results demonstrate that the proposed method can significantly increase the FIB observable rate under 100% routability.

This paper presents a novel method for functional test generation aiming at exploring control state space of the design. The steady-state probabilities (SP’s) of the abstract design's control FSM are used to guide test generation. The SP's of the states can reflect how hard the states can be reached, and the hard-to-reach states have high priority to be exercised. Experimental results show that our method has better performance in test generation in comparison with constrained random simulation, and demonstrate that SP's provide good guidance on traversing hard-to-reach states of the design under validation.

Keywords - functional test generation; steady-state probability; control state space exploration

Testing on system level using a static and homogeneous architecture of clients is common practice. This paper introduces a new approach to use a heterogeneous and dynamic system of resource restricted test clients for automated testing. Due to changing resources and availability of the clients, the test case distribution needs to be recalculated dynamically during the test execution. All necessary conditions and parameters are represented by a formal model. It is shown that the algorithmic problem of DYNAMIC TESTPARTITIONNING can be solved in polynomial time by a heuristic recursive algorithm. A testbench architecture is introduced and by simulation it is shown that the testbench can execute the test requirements within a small variation using a number of several hundred clients. The simulation can react dynamically on changing resources and availability of the test clients within several seconds. The approach is generic and can be adapted to a huge number of systems.

11.5 Memory Resource Allocation and Scheduling in MPSoC

Chairs: Andreas Herkersdorf, Technische Universität München, DE; Donatella Sciuto, Politecnico di Milano, IT

Integrating Multi-Processor System-on-Chips (MPSoC) with 3D-stacked reconfigurable SRAMs.

With the availability of advanced MPSoC and emerging Dynamic RAM (DRAM) interface technologies, an optimal allocation of logical data buffers to physical memory cannot be handled manually anymore due to the huge design space. An allocation does not only need to decide between an on-chip FIB probing, but also needs to take an increasing number of available memory channels, different bandwidth capacities and various routing possibilities into account. We formalize this problem and introduce a Mixed Integer Linear Programming (MILP) model based on two different optimization criteria. We implement the MILP model into a retargetable tool and present a case study with representative data of the Long-Term Evolution (LTE) standard to show the real-life applicability of our approach.

Synchronous dataflow graphs (SDFGs) are widely used to model digital signal processing (DSP) and streaming media applications. In this paper, we use retiming to optimize SDFGs to achieve a high throughput with low storage requirement. Using a memory constraint as an additional enabling condition, we define a memory constrained self-timed execution of an SDFG. Exploiting the state-space generated by the execution, we can check whether a retiming exists that leads to a rate-optimal schedule under the memory constraint. Combining this with a binary search strategy, we present a heuristic method to find a proper retiming and a static scheduling which schedules the retimed SDFG with optimal rate (i.e., maximal throughput) and with as little storage space as possible. Our experiments are carried out on hundreds of synthetic SDFGs and several models of real applications. Deterministic graph results and real application results show that, in 76% of the tested models, our method leads to a retimed SDFG whose rate-optimal schedule requires less storage space than the proven minimal storage requirement of the original graph, and in 20% of the cases, the returned storage requirements equal the minimal ones. The average improvement is about 7.3%. The results also show that our method is computationally efficient.

A Constraint-Based Design Space Exploration Framework for Real-Time Applications on MPSoCs

Kathrin Rovsavl, Ingo Sander

Design space exploration (DSE) is a critical step in the design process of real-time multiprocessor systems. Combining a formal base in form of SDF graphs with predictable platforms providing guaranteed QoS, the paper proposes a flexible and extendable DSE framework that can provide performance guarantees for multiple applications implemented on a shared platform. The DSE framework is formulated in a declarative style as interprocess communication-aware constraint programming (CP) model. Apart from mapping and scheduling of
application graphs, the model supports design constraints on several cost and performance metrics, as e.g. memory consumption and achievable throughput. Using constraints with different compliance level, the framework introduces support for mixed criticality in the CP model. The potential of the approach is demonstrated by means of experiments using a Sobel filter, a SUSAN filter, a RASTAPLP application and a JPEG encoder.

Reliability-Aware Mapping Optimization of Multi-Core Systems with Mixed-Criticality
Shin-Haeng Kang, Hoeseok Yang, Sungchan Kim, Iuliana Backovar, Soonhoi Ha, Lothar Thiele
This paper presents a novel mapping optimization technique for mixed critical multi-core systems with different reliability requirements. For this scope, we derived a quantitative reliability metric and presented a scheduling analysis that certifies given mixed-criticality constraints. Our framework is capable of investigating re-execution, passive replication, and modular redundancy with optimized voter placement, while typical hardening approaches consider only one or two of these techniques. The proposed technique complies with existing safety standards and is power-efficient, as demonstrated by our experiments.

From Simulink to NoC-based MPSoC on FPGA
Francesco Robino, Johnny Öberg
Network-on-chip (NoC) based multi-processor systems are promising candidates for future embedded system platforms. However, because of their complexity, new high level modeling techniques are needed to design, simulate and synthesize embedded systems targeting NoC-based MPSoC. Simulink is a popular modeling environment suitable to model at system level. However, there is no clear standard to synthesize Simulink models into SW and HW towards a NoC-based MPSoC implementation. In addition, many of the proposed solutions require large overhead in terms of SW components and memory requirements, resulting in complex and customized multi-processor platforms. In this paper we present a novel design-flow to synthesize Simulink models onto a NoC-based MPSoC running on low-cost FPGAs. Our design flow constrains the MPSoC and the Simulink model to share a common semantics domain. This permits to reduce the need of resource consuming SW components, reducing the memory requirements on the platform. At the same time, performances (throughput) of dataflow applications can be increased when the number of processors of the target platform is increased. This is shown through a case study on FPGA.

11.6 System-Level Thermal Estimation and Management
Chair: Coskun Ayse, Boston University, US; Wolfgang Nebel, OFFIS, DE

Minimal Sparse Observability of Complex Networks: Application to MPSoC Sensor Placement and Run-time Thermal Estimation & Tracking
Sanatun Samsa, Nikil Dutt
This paper addresses the fundamental and practically useful question of identifying a minimum set of sensors and their locations through which a large complex dynamical network system and its time-dependent states can be observed. The paper defines the minimal sparse observability problem (MSOP) and provides analytical tools with necessary and sufficient conditions to make an arbitrary complex dynamical network system completely observable. The mathematical tools are then used to develop effective algorithms to find the sparsest measurement vector that provides the ability to estimate the internal states of a complex dynamic network system from experimentally accessible outputs. The developed algorithms are further used in the design of a sparse Kalman filter (SKF) to estimate the time-dependent internal states of a linear time-invariant (LTI) dynamical network system. The approach is applied to illustrate the minimization process.

mDTM: Multi-Objective Dynamic Thermal Management for On-Chip Systems
Helia Khadr, Thomas Ebi, Muhammad Shailique, Hussam Artsouh, Jörg Henkel
Thermal hot spots and unbalanced temperatures between chips on a chip can cause either degradation in performance or may have a severe impact on reliability, or both. In this paper, we propose mDTM, a proactive dynamic thermal management technique for on-chip systems. It employs multi-objective management for migrating tasks in order to both prevent the system from hitting an undesirable thermal threshold and to balance the temperatures between the cores. Our evaluation on the Intel SCC platform shows that mDTM can successfully avoid a given thermal threshold and reduce spatial thermal variation by 22%. Compared to state-of-the-art, our mDTM achieves up to 58% performance gain. Additionally, we deploy an FPGA and IR camera based setup to analyze the effectiveness of our technique.

Thermal Management of Batteries Using a Hybrid Supercapacitor Architecture
Donghwa Shin, Massimo Poncon, Enrico Macii
Thermal analysis and management of batteries have been an important research issue for battery-operated systems such as electric vehicles and mobile devices. Nowadays, battery packs are designed considering heat dissipation, and external cooling devices such as a cooling fan are also widely used to enhance the reliability and extend the lifetime of a battery. This type of approaches that target the enhancement of the cooling efficiency via the reduction of the thermal resistance cannot achieve an immediate temperature drop to avoid a thermal emergency situation. Approaches based on removing the heat from the heat sources via idle period insertion (similar to what is done for silicon devices) would allow faster thermal response; however, it is not obvious how to implement these schemes in the context of batteries. In this paper, we propose the use of a simple parallel battery-supercapacitor hybrid architecture with a dual-mode discharging strategy that can provide immediate temperature management, in which the supercapacitor is used as an energy buffer during the idle periods of the battery. Simulation results show that the proposed method can keep the battery temperature within the safe range without external cooling devices while exploiting the advantages of the battery-supercapacitor parallel connection.

Thermal Analysis and Model Identification Techniques for a Logic + WIDEIO Stacked DRAM Test Chip
Francesco Beneventi, Andrea Bartolini, Pascal Vivet, Denis Dutot, Luca Benini
High temperature is one of the limiting factors and major concerns in 3D-chip integration. In this paper we use a 3D test chip (WIDEIO DRAM on top of a logic die) equipped with temperature sensors and heaters to explore thermal effects. We correlated real thermal measurements with the power dissipated by the heaters using model learning techniques. The resulting compact thermal model is able to predict temperatures at chip locations far from the temperature sensors and to infer the power dissipation at any location of the chip. Results are verified by mean of an off-sample validation technique and show a high accuracy of the compact thermal model when compared with silicon measurements.

Adaptive Power Allocation for Many-Core Systems Inspired from Multicagent Auction Model
Xiaohang Wang, Baixin Zhao, Terrence Mak, Mei Yang, Yingtao Jiang, Masoud Daneshstabl, Maurizio Palesi
Scaling of future many-core chips is hindered by the challenge imposed by ever-escalating power consumption. At its worst, an increasing fraction of the chips will have to be shut down, as power supply is inadequate to simultaneously switch all the transistors. This so-called dark silicon problem brings up a critical issue regarding how to achieve the maximum performance within a given limited power budget. This issue is further complicated by two facts. First, high variation in power budget calls for wide range power control capability, whereas most current frequency/voltage scaling techniques cannot effectively adjust power over such a wide range. Second, as the applications’ behavior becomes more complicated, there is a pressing need for scalability and global coordination, rendering heuristic-based centralized or fully distributed control schemes inefficient. To address the aforementioned problems, in this paper, a power allocation method employing multicagent auction models is proposed, referred as Hierarchical MultiAgent based Power allocation (HiMAP). Ties act the role of consumers to bid for power budget and the whole process is modeled by a combinatorial auction, whereas HiMAP finds the Walrasian equilibria. Experimental results have confirmed that HiMAP can reduce the execution time by as much as 45% compared to three competing methods. The runtime overhead and cost of HiMAP are also small, which makes it suitable for adaptive power allocation in many-core systems.

Keywords - many-core; power allocation; multicagent...
11.7 Power and Emerging Technologies in Reconfigurable Computing

Chairs: Diana Goehringer, Ruhr-Universität Bochum (RUB), DE; Fabrizio Ferrandi, Politecnico di Milano, IT

Exploiting STT-NV Technology for Reconfigurable, High Performance, Low Power, and Low Temperature Functional Unit Design
Ashrat Reddy Ashammaragari, Hamid Mahmoodi, Herouari Komayoun

Unavailability of functional units and their unequal activity makes performance bottlenecks and thermal hot spots in general-purpose processors. We propose to use reconfigurable functional units to overcome these challenges. A selected set of complex functional units that might be underutilized, such as a multiplier and divider, are realized in a time-multiplexed fashion using a shared programmable Look Up Table (LUT) based fabric. This allows for run-time reconfiguration and migration of their activity. LUT-based implementation also allows under-utilized functional units to be dynamically reconfigured to the functional units that have a performance bottleneck and hence improving performance. The programmable LUTs are realized using Spin Transfer Torque (STT) Magnetic technology (also called STT-NV) due to its zero leakage and CMOS compatibility. The results show significant performance improvement of 16% on average across standard benchmarks, when replacing CMOS multiplier and divider with reconfigurable STT-NV LUT counterpart. In addition, reconfiguration reduces the maximum temperature of functional units by up to 27°C and almost eliminates the thermal variation across them. This comes with small power overhead and no area impact.
Keywords--STT-NV logic; reconfigurable architecture; low power; functional units; low temperature; multiplier; divider

A Novel Embedded System for Vision Tracking
Antonis Nikitakis, Theofilos Paganos, Ioannis Papaefstathiou

One of the most important challenges in the field of Computer Vision is the implementation of low-power embedded systems that will execute very accurate, yet real-time, algorithms. In this paper, we present a novel embedded system implementing this algorithm. We accelerate the bottleneck of the algorithm by designing and implementing a high bandwidth distributed memory sub-system which is independent of the various software parameters. We demonstrate the applicability and efficiency of this novel approach by implementing our scheme in a modern FPGA.
Keywords - Embedded System, Object Tracking, Distributed Memory, Random Forests, Classification, FPGA.

11.8 Embedded Tutorial: GPGPUs: How to Combine High Computational Power with High Reliability

Organizer: Matteo Sonza Reorda, Politecnico di Torino, IT
Chairs: Dimitis Gizopoulos, University of Athens, GR; Rob Atken, ARM, US

GPGPUs are used increasingly in several domains, from gaming to different kinds of computationally intensive applications. In many applications GPGPU reliability is becoming a serious concern. In this tutorial, we will discuss techniques to improve the reliability of GPGPUs. We will cover topics such as hardware redundancy, error-correcting codes, and software-based reliability enhancement strategies. By the end of the tutorial, participants should have a good understanding of the state-of-the-art in GPGPU reliability and be able to apply these techniques to their own projects.
Keywords - GPGPU, reliability, error correction, hardware redundancy, software-based reliability.
issue, and several research activities are focusing on its evaluation. This paper offers an overview of some major results in the area. First, it shows and analyzes the results of some experiments assessing GPGPU reliability in HPC datacenters. Second, it provides some recent results derived from radiation experiments about the reliability of GPGPUs. Third, it describes the characteristics of an advanced fault-injection environment, allowing effective evaluation of the resiliency of applications running on GPGPUs.

Keywords - GPGPUs, reliability, HPC, fault injection, radiation experiments

12.1 Hot Topic: The Future of Interfacing to the Natural World

Organizers: Ian O'Connor, Lyon Institute of Nanotechnology, FR; Thomas Mikolajick, NamLab gGmbH, DE

Chairs: Thomas Mikolajick, NamLab gGmbH, DE; Ian O'Connor, Lyon Institute of Nanotechnology, FR

[350] Integrated Circuits Processing Chemical Information: Prospects and Challenges
A. Richter, A. Voigt, R. Schütz, S. Henker, M. Volp

The unbelievable properties of our information processing capabilities regarding the processing of big data, resilience, and energy efficiency are inspiration sources for the optimization and the rethinking of the principles of electronic information processing. Here, we present an approach of integrated circuits intended to solve chemical problems by active processing of chemical information.

[351] Interfacing to Living Cells
Rudy Lauwereins

Recent advances in More than Moore technology enable close inspection of and even direct interfacing to living cells. This paper illustrates this through three use cases. In the first use case, the type or quality of billions of cells is quickly inspected in a fluidic medium. Secondly, the effect of potential drugs is monitored in neural cell cultures. In the third use case, neural brain activity is recorded in vivo using implantable electrodes to understand how the brain functions.

Keywords - interfacing, cells, circulating tumor cells, stem cell cultures, lens free imaging, micro-fluidics, neuroprobes

[352] Video Analytics Using Beyond CMOS Devices
Vijaykrishnan Narayanan, Suman Datta, Gert Cauwenberghs, Don Chiualli, Steve Levitan, Philip Wong

The human vision system understands and interprets complex scenes for a variety of visual tasks in real-time while consuming less than 20 Watts of power. The holistic design of artificial vision systems will approach and eventually exceed the capabilities of human vision systems by a grand challenge. The design of such a system needs advances in multiple disciplines. This paper focuses on advances needed in the computational fabric and provides an overview of a new-genre of architectures inspired by advances in both the understanding of the visual cortex and the emergence of devices with new mechanisms for state computations.

Keywords - Emerging Device; Coupled Oscillator

[353] Energy Efficient Neural Networks for Big Data Analytics
Yu Wang, Boxun Li, Rong Luo, Yuan Chen, Ningyi Xu, Huazhong Yang

The world is experiencing a data revolution to discover knowledge in big data. Large scale neural networks are one of the mainstream tools of big data analytics. Processing big data with large scale neural networks includes two phases: the training phase and the operation phase. Huge computing power is required to support the training phase. And the energy efficiency (power efficiency) is one of the major considerations of the operation phase. We first explore the computing power of GPGPUs for big data analytics and demonstrate an efficient GPU implementation of the training phase of large scale recurrent neural networks (RNNs). We then introduce a promising ultrahigh energy efficient implementation of neural networks’ operation phase by taking advantage of the emerging memristor technique. Experiment results show that the proposed GPU implementation of RNNs is able to achieve 2~11x speed-up compared with the basic GPU implementation. And the scaled-up recurrent neural network trained with GPUs realizes an accuracy of 47% on the Microsoft Research Sentence Completion Challenge, the best result achieved by a single RNN on the same dataset. In addition, the proposed memristor-based implementation of neural networks demonstrates power efficiency of > 400 GFLOPS/W and achieves energy savings of 22x on the HMAX model compared with its pure digital implementation counterpart.

[354] Special Session: How Secure are PUFs Really? On the Reach and Limits of Recent PUF Attacks
Ulrich Rührmair, Ulf Schlichtmann, Wayne Burleson

Just over a decade ago, Physical Unclonable Functions (PUFs) have been introduced as a new cryptographic and security primitive in a number of seminal publications. Due to their assumed security and cost advantages, they have attracted substantial attention both from the security industry and the academic community, and are also gaining ground in commercial applications. Nevertheless, a number of recent works have presented successful attacks on PUF core properties, such as their digital and physical unclonability. How strong and relevant are these attacks, and how secure are PUFs really? This question is addressed in a dedicated hot topic session at DATE 2014. This paper provides a short and easily accessible overview of the session.

Index Terms - Physical Unclonable Functions (PUFs), Weak PUFs, Strong PUFs, Security, Modeling Attacks, Invasive Attacks, Side Channel Attacks, Protocol Attacks

[355] PUFs at a Glance
Ulrich Rührmair, Daniel E. Holcomb

Physical Unclonable Functions (PUFs) are a new, hardware-based security primitive, which has been introduced just about a decade ago. In this paper, we provide a brief and easily accessible overview of the area. We describe the typical security features, implementations, attacks, protocols uses, and applications of PUFs. Special focus is placed on the two most prominent PUF types, so-called “Weak PUFs” and “Strong PUFs”, and their mutual differences.

Keywords - Physical Unclonable Functions, Overview, Survey, Weak PUFs, Strong PUFs

[356] PUF Modeling Attacks: An Introduction and Overview
Ulrich Rührmair, Jan Stüber

Machine learning (ML) based modeling attacks are the currently most relevant and effective attack form for so-called Strong Physical Unclonable Functions (Strong PUFs). We provide an overview of this method in this paper: We discuss (i) the basic conditions under which it is applicable; (ii) the ML algorithms that have been used in this context; (iii) the latest and most advanced results; (iv) the right interpretation of existing results; and (v) possible future research directions.

Index Terms - Physical Unclonable Functions, Machine Learning, Modeling Attacks, Cryptanalysis

[357] Hybrid Side-Channel / Machine-Learning Attacks on PUFs: A New Threat?
Xiaolin Xu, Wayne Burleson

Machine Learning (ML) is a well-studied strategy in modeling Physical Unclonable Functions (PUFs) but reaches its limits while applied on instances of high complexity. To address this issue, side-channel attack is combined to help reduce the computational workload of ML modeling attacks and make it more applicable. In this work, we present the currently known hybrid
side-channel attacks on PUFs. A taxonomy is proposed based on the characteristics of different side-channel attacks. The practical reach of some published side-channel attacks is discussed. Both challenges and opportunities for PUF attackers are introduced. Countermeasures against some certain sidechannel attacks are also analyzed. To better understand the side-channel attacks on PUFs, three different methodologies of implementing side-channel attacks are compared. At the end of this paper, we bring forward some open problems for this research area.

[358] Physical Vulnerabilities of Physically Undeniable Functions
Clemens Heilmeier, Christian Bold, Dmitriy Nedosposav, Shahin Tajik, Jean-Pierre Seillier
In recent years one of the most popular areas of research in hardware security has been physically undeniable functions (PUF). PUFs provide primitives for implementing tamper detection, encryption and device fingerprinting. One particularly common application is replacing non-volatile memory (NVM) as key storage in embedded devices like smart cards and secure microcontrollers. Though a wide array of PUF have been proposed in the academic literature, vendors have only begun to roll out PUFs in their end-user products. Moreover, the improvement to overall system security provided by PUFs is still the subject of much debate. This work reviews the state of the art of PUFs in general, and as a replacement for key storage in particular. We review also techniques and methodologies which make the physical response characterization and physical/digital cloning of PUFs possible.

[359] Protocol Attacks on Advanced PUF Protocols and Countermeasures
Martijn van Dijk, Ulrich Rührmair
In recent years, PUF-based schemes have not only been specified for the basic security tasks of tamper sensitive key storage or system identification, but also for more complex cryptographic protocols like oblivious transfer (OT), bit commitment (BC), or key exchange (KE). These more complex protocols are secure against adversaries in the stand-alone, good PUF model. In this survey, a shortened version of [17], we explain the stronger bad PUF model and PUF re-use model. We argue why these stronger attack models are realistic, and that existing protocols, if used in practice, will need to face these. One consequence is that the design of advanced cryptographic PUF protocols needs to be strongly reconsidered. It suggests that strong PUFs require additional hardware properties in order to be broadly usable in such protocols; Firstly, they should ideally be erasable, making that single PUF-response can be erased without affecting other responses. If the area efficient implementation of this feature turns out to be difficult, new forms of Controlled PUFs [3] (such as Logically Erasable and Logically Reconfigurable PUFs [6]) may suffice in certain applications. Secondly, PUFs should be certifiable, meaning that one can verify that the PUF has been produced faithfully and has not been manipulated in any way afterwards. The combined implementation of these features represents a pressing and challenging problem for the PUF hardware community.

Index Terms - (Strong) Physical Undeniable Functions; (Strong) PUFs; Attack Models; Oblivious Transfer; Bit Commitment; Key Exchange; Erasable PUFs

Masoud Rostami, James B. Wendt, Mostag Polkong, Farinaz Koushanfar
The physical undeniable function (PUF) has emerged as a popular and widely studied security primitive based on the randomness of the underlying physical medium. To date, most of the research emphasis has been placed on finding new ways to measure randomness, hardware realization and analysis of a few initially proposed structures, and conventional secretkey based protocols. In this paper, we present our recent analysis of the emerging security and future trends in this area that aim to change the scope, widen the application domain, and make a lasting impact. We emphasize on the development of new PUF-based primitives and paradigms, robust protocols, publickey protocols, digital PUFs, new technologies, implementations, metrics and tests for evaluation/verification, as well as relevant attacks and countermeasures.

12.3 Multimedia Systems

Chair: Theochanides Theochanis, University of Cyprus, CY; Cristiana Bokthini, Politecnico di Milano, IT

[361] Flexible and Scalable Implementation of H.264/AVC Encoder for Multiple Resolutions Using ASIPs
Hong Chinh Doan, Hari Javaid, Sri Parameswaran
Real-time encoding of video streams is computationally intensive and rarely carried out at high resolutions. In this paper, for the first time, we propose a platform for H.264 encoder which is both flexible (allows software upgrades and scalable (supports multiple resolutions), and supports high video quality (by using both intraprediction and interprediction) and high throughput (by exploiting slice-level and pixel-level parallelisms). Our platform uses multiple Application Specific Instruction Set Processors (ASIPs) with local shared memories, and hardware accelerators (in the form of custom instruction sets). Our platform can be configured to use a particular number of ASIPs (slices per video frame) for a specific video resolution at design-time. The MPSoC architecture is automatically generated by our platform and the H.264 software does not need any modification, which enables quick design space exploration. We implemented the proposed platform in a commercial design environment, and illustrated its utility by creating systems with up to 170 ASIPs supporting resolutions up to HD1080. We further show how power gating can be used in our platform to save energy consumption.

[362] A Flexible ASIP Architecture for Connected Components Labeling in Embedded Vision Applications
Juan Fernando Eusse, Rainer Leupers, Gerd Ascheid, Patrick Sudowe, Bastian Leibe, Taron Sadasue
Real-time identification of connected regions of pixels in large (e.g. FullHD) frames is a mandatory and expensive step in many computer vision applications that are becoming increasingly popular in embedded mobile devices such as smartphones, tablets and head mounted devices. Standard off-the-shelf embedded processors are not yet able to cope with the performance/flexibility trade-offs required by such applications. Therefore, in this work we present an Application Specific Instruction Set Processor (ASIP) tailored to concurrently execute thresholding, connected components labeling and basic feature extraction of image frames. The proposed architecture is capable to cope with frame complexities ranging from QCIF to FullHD frames with 1 to 4 bytes-per-pixel formats, while achieving an average frame rate of 30 frames-per-second (fps). Synthesis was performed for a standard 65nm CMOS library, obtaining an operating frequency of 350MHz and 2.1mm2 area. Moreover, evaluations were conducted both on typical and synthetic data sets, in order to thoroughly assess the achievable performance. Finally, an entire planar-marker based augmented reality application was developed and simulated for the ASIP.

[363] Image Progressive Acquisition for Hardware Systems
Janxiong Liu, Chistos Bouganim, Peter V.K. Cheung
As the resolution of digital images increases, accessing raw image data from memory has become a major consideration during the design of image/video processing systems. This is due to the fact that the bandwidth requirement and energy consumption of such image accessing process has increased. Inspired by the successful application of progressive image sampling techniques in many image processing tasks, this work proposes to apply similar concept within hardware systems to efficiently trade image quality for reduced memory bandwidth requirement and lower energy consumption. Based on this idea, a hardware system is proposed that is placed between the memory subsystem and the processing core of the design. The proposed system alters the conventional memory access pattern to progressively and adaptively access pixels from a target memory external to the system. The sampled pixels are used to reconstruct an approximation to the ground truth, which is stored in an internal image buffer for further processing. The system is prototyped on FPGA and its performance evaluation shows that a saving of up to 85% of memory accessing time and 33%/45% of image acquisition time/energy is achieved on the benchmark image "lena" while maintaining a PSNR of about 30 dB.

[364] High-Quality Real-Time Hardware Stereo Matching Based on Guided Image Filtering
Chistos Ttofis, Theocharis Theocharides
Stereo matching is a vital task in several emerging embedded vision applications requiring high-quality depth computation and real-time frame-rate. Although several stereo matching dedicated hardware systems have been proposed in recent years, only few of them focus on balancing accuracy and speed. This paper proposes a hardware-based stereo matching architecture that aims to provide high accuracy and concurrently high performance in embedded vision applications. The proposed architecture integrates a compact and efficient design of the recently proposed guided image filter; an edge-preserving filter that reduces the hardware complexity of the implemented stereo algorithm, while at the same time maintains high-quality results. A prototype of the architecture has been implemented on a Xilinx-7 FPGA board, achieving 60 fps for 720p resolution images. Moreover, the proposed design delivers leading accuracy when compared to state-of-the-art hardware implementations.
12.4 Physical Aspects

Chair: Carl Sechen, University of Texas at Dallas, US; Jens Lieng, TU Dresden, DE

[365] Optimization of Standard Cell Based Detailed Placement for 16 nm FinFET Process
Yue Lin, Martin D. F. Wong

FinFET transistors have great advantages over traditional planar MOSFET transistors in high performance and low power applications. Major foundries are adopting the FinFET technology for CMOS semiconductor device fabrication in the 16 nm technology node and beyond. Edge device degradation is among the major challenges for the FinFET process. To avoid such degradation, dummy gates are needed on device edges, and the dummy gates have to be tied to power rails in order not to introduce unconnected parasitic transistors. This requires that each dummy gate must abut at least one source node after standard cell placement. If the drain nodes at two adjacent cell boundaries abut each other, additional source nodes must be inserted in between for dummy gate power tying, which costs more placement area. Usually there is some flexibility during detailed placement to horizontally flip the cells or switch the positions of adjacent cells, which has little impact on the global placement objectives, such as timing conditions and net congestion. This paper proposes a detailed placement optimization strategy for the standard cell-based designs. By flipping a subset of cells in a standard cell row and switching pairs of adjacent cells, the number of drain to drain abutments between adjacent cell boundaries can be optimally minimized, which saves additional source node insertion and reduces the length of the standard cell row. In addition, the proposed graph model can be easily modified to consider more complicated design rules. The experimental results show that the optimization of 100k cells is completed within 0.1 second, verifying the efficiency of the proposed algorithm.

[366] Signature Indexing of Design Layouts for Hotspot Detection
Christian Andrades, M. Andrea Rodriguez, Charles C. Chiang

This work presents a new signature for 2D spatial configurations that is useful for the optimization of a hotspot detection process. The signature is a string of numbers representing changes along the horizontal and vertical slices of a configuration, which serves as the key of an inverted index that groups layout windows with the same signature. The method extracts signatures from a compact specification of similar exact patterns with a fixed size. Then, these signatures are used as search keys of the inverted index to retrieve candidate windows that can match the patterns. Experimental results show that this simple type of signature has 100% recall and, in average, over 85% of precision in terms of the area effectively covered by the pattern and the retrieved area of the layout. In addition, the signature shows a good discriminate quality, since around 99% of the extracted signatures match each of them with a single pattern.

[367] Metal Layer Planning for Silicon Interposers with Consideration of Routability and Manufacturing Cost
Wen-Hao Liu, Tzu-Kai Chien, Ting-Chi Wang

A 2.5D IC provides a silicon interposer to integrate multiple dies into a package, which not only offers better performance than 2D ICs but also has lower manufacturing complexity than true 3D ICs. In an interposer, routing wires connect signals between dies or route signals from dies to the package substrate. The number of metal layers in an interposer is one of the critical factors to affect the routability and manufacturing cost of the 2.5D IC. Thus, how to achieve 100% routing completion rate in an interposer using a minimum number of metal layers plays a key role for the success of a 2.5D IC. This paper presents a global-routing-based metal layer planner called VGR to identify a minimal number of metal layers for an interposer with consideration of routability and manufacturing cost. Also, VGR can identify a good stacking order of the horizontal and vertical layers in an interposer such that the routing solution in the interposer costs fewer tiers. To our best knowledge, this paper is the first study to solve the metal layer planning problem for silicon interposers.

12.5 System-level Design Space Exploration

Chair: Federico Petro, TIMA, FR; Luciano Lavagno, Politecnico di Torino, IT

[368] Non-Intrusive Integration of Advanced Diagnosis Features in Automotive E/E-Architectures
Ulrich Abelein, Aleksandro Cook, Piet Engeke, Michael Glaß, Felix Reimann, Laura Rodriguez Gómez, Thomas Russ, Jürgen Teich, Dominik Ut, Hans-Joachim Wunderlich

With ever more complex automotive systems, the current approach of using functional tests to locate faulty components results in very long analysis procedures and poor diagnostic accuracy. Built-In Self-Test (BIST) offers a promising alternative to collect structural diagnostic information during E/E-architecture test. However, as the automotive industry is quite cost-driven, structural diagnosis shall not deteriorate traditional design objectives. With this goal in mind, the work at hand proposes a design space exploration to integrate structural-diagnostic capabilities into an E/E-architecture design. The proposed integration is performed non-intrusively, i.e., the addition and execution of tests (a) does not affect any functional applications and (b) does not require any costly changes in the communication schedules.

Kumud Nepal, Yueying Li, R. Iris Bahar, Shenef Reda

Many classes of applications, especially in the domains of signal and image processing, computer graphics, computer vision, and machine learning, are inherently tolerant to inaccuracies in their underlying computations. This tolerance can be exploited to design approximate circuits that perform within acceptable accuracies but have much lower power consumption and smaller area footprints (and often better run times) than their exact counterparts. In this paper, we propose a new class of automated synthesis methods for generating approximate circuits directly from behavioral-level descriptions. In contrast to previous methods that operate at the Boolean level or use custom modifications, our automated behavioral synthesis method enables a wider range of possible approximations and can operate on arbitrary designs. Our method first creates an abstract synthesis tree (AST) from the input behavioral description, and then applies variant operators to the AST using an iterative stochastic greedy approach to identify the optimal inexact designs in an efficient way. Our method is able to identify the optimal designs that represent the Pareto frontier trade-off between accuracy and power consumption. Our methodology is developed into a tool we call ABACUS, which we integrate with a standard ASIC experimental flow based on industrial tools. We validate our methods on three realistic Verilog-based benchmarks from three different domains - signal processing, computer vision, and machine learning. Our tool automatically discovers optimal designs, providing area and power savings of up to 55% while maintaining good accuracy.

[370] Automatic Generation of Custom SIMD Instructions for Superword Level Parallelism
Taemin Kim, Yatir Hekote

Application specific instruction-set processors (ASIPs) have drawn significant attention from System-on-a-Chip (SoC) community due to the capability of fine grain flexibility and customizability. In order to maximize the benefit of ASIP, automatic instruction set extension (ISE) is required. In the past decade, there have been plethora of researches on automatic ISE for custom scalar instruction. However, due to increasing usage of SIMD instructions to exploit data level parallelism (DLP) that exists both across loop iterations and within a basic block called Superword Level Parallelism (SLP), automatic generation of custom SIMD instructions is the inevitable direction of automatic ISE. In this paper, we propose an algorithm that automatically generates custom SIMD instructions from a set of custom scalar instructions to exploit SLP. We have demonstrated 52.4% and 30.8% performance improvement on average over base instruction set and additional custom scalar instructions, respectively.

Emanuele Cannella, Mohamed A. Bamakhrama, Todor Stefanov

Modern multiprocessor streaming systems have hard real-time constraints that must be always met to ensure correct functionality. At the same time, these streaming systems must be designed to use the minimum required amount of resources (such as processors and memory). In order to meet such constraints, using scheduling algorithms from the classical real-time scheduling theory represents an attractive solution approach. These algorithms enable: (1) providing timing guarantees to the applications running on the system, and (2) deriving analytically the minimum number of processors required to schedule the applications. So far, designers in the embedded systems community have focused on global and partitioned

Keywords - Stereo Matching; Embedded Systems; FPGAs
12.6 Error Resilience and Power Management

Chairs: William Fomicaci, Politecnico di Milano; DEIB, IT; Kim Guehnetter, OFFIS, DE

Many applications produce acceptable results even when their underlying computations are executed in an approximate manner. For such applications, approximate circuits enable hardware implementations that exhibit improved efficiency for a given quality. Previous efforts have largely focused on the design of approximate combinational logic blocks such as adders and multipliers. In practise, however, designers are concerned with the quality of outputs generated by a sequential circuit after several cycles of computation, rather than an embedded combinational block. We propose ASLAN (Automatic methodology for Sequential Logic Approximation), the first effort towards the synthesis of approximate sequential circuits. Given a sequential circuit and an output quality constraint, ASLAN creates an approximate version of the circuit that consumes lower energy, while meeting the specified quality bound. The key challenges in approximating sequential circuits are (i) to model how errors due to approximations are generated, re-circulate through the combinational logic over multiple cycles of operation, and eventually impact quality of the final output, and (ii) to select the most beneficial approximations, i.e., those that result in higher energy savings for smaller impact on quality. ASLAN addresses the first challenge by constructing a virtual Sequential Quality Constraint Circuit (SQCC) and utilizing formal verification techniques to ensure that the selected approximations meet the quality constraint. To address the second challenge, ASLAN identifies combinational blocks in the sequential circuit that are amenable to approximation, generates local quality-energy trade-off curves for them, and uses a gradient-descent approach to iteratively approximate the entire sequential circuit. We used ASLAN to automatically synthesize approximate versions of ten sequential benchmarks, resulting in energy reductions of 1.20×-2.44× for tight quality constraints, and 1.32×-4.42× for moderate quality constraints. We present case studies of using the approximate circuits generated by ASLAN in two popular applications: MPEG Encoding and K-Means Clustering - obtaining 1.32X energy savings with 0.5% PSNR degradation, and 1.20X energy savings with 0.8% increase in mean cluster radius, respectively.

Index Terms - Low-power Design, Approximate Computing, Approximate Circuits, Logic Synthesis, Sequential circuits

12.7 Built-In self-test Solutions for Mixed-signal and RF ICs

Chairs: Jacob A. Abraham, University of Texas at Austin; US; Marian Verhelst, KU Leuven, BE

This paper presents a Built-in self-test (BIST) solution for polar transmitters with low cost. Polar transmitters are desirable for portable devices due to higher power efficiency they provide overall learning capability of our chip. The BIST-related effectiveness is then evaluated on two case studies: the detection of parametric and catastrophic faults in an LNA and an RF front-end circuit, time borrowing can be applied between error checkpoints thereby avoiding bubbles which could occur in the existing Bubble Razor design with a checkpoint at every latch on the critical path. Significant overhead in performance and power consumption due to too frequent occurrence of bubble generations. We present a design methodology for coarse-grained Bubble Razor margin, Bubble Razor was introduced to dynamically detect and correct errors in two-phase transparent latch designs [13]. However, it does not fully exploit the potential of two-phase transparent latch design, e.g. time borrowing. Thus, especially at low supply voltage where the effect of process variation becomes significant, the existing Bubble Razor can suffer from significant overhead in performance and power consumption due to too frequent occurrence of bubble generations. We present a design methodology for coarse-grained Bubble Razor which exploits the time-borrowing characteristic of two-phase transparent latch design. By selectively inserting error checkpoints, i.e., shadow latches and error management logic, in the circuit, time borrowing can be applied between error checkpoints thereby avoiding bubbles which could occur in the existing Bubble Razor design with a checkpoint at every latch on the critical path. We present a methodology to choose the gain size (the number of stages between error checkpoints) based on 3-sigma delay distribution. We also verify the benefits of coarse-grained Bubble Razor with a real microprocessor, Core-A design [15] using 20nm Predictive Technology Model (PTM) [16]. The proposed methodology offers 62% improvement in performance (MIPS) and 49% less energy consumption (per instruction) at 0.6V operation (zero frequency margin) over the original Bubble Razor scheme. In addition, it gives 25% area reduction in core design.

This paper presents a Built-in self-test (BIST) solution for polar transmitters with low cost. Polar transmitters are desirable for portable devices due to higher power efficiency they provide overall learning capability of our chip. The BIST-related effectiveness is then evaluated on two case studies: the detection of parametric and catastrophic faults in an LNA and an RF front-end circuit, respectively.

Index Terms - Low-power Design, DC-DC converter, Power delivery network, Multicore, Consolidation.

This paper introduces a novel sensor-less, event-driven power analysis framework called FEPMA for providing highly accurate and nearly instantaneous estimates of power dissipation in an Android smartphone. The key idea is to collect and correctly record various events of interest within a smartphone as applications are running on the application processor within it. This is done in turn by instrumenting the Android operating system to provide information about power/performance state changes of various smartphone components at the lowest layer of the kernel to avoid time stamping delays and component state observability issues. This technique then enables one to perform fine-grained (in time and space) power metering in the smartphone. Experimental results show significant improvement compared to previous approaches and good fidelity with respect to actual current measurements. The estimation error of the proposed method is lower by a factor of two than the state-of-the-art method.
compared to traditional Cartesian transmitters. However, they generally require iterative test/measurement/calibration cycles. The delay skew between the envelope and phase signals and the finite envelope bandwidth can create intermodulation distortion (IMD) that leads to the violation of the spectral mask and error vector magnitude (EVM) requirements. Typically, these parameters are not directly measured but calibrated through spectral performance analysis using expensive RF equipment, leading to lengthy and costly measurement/calibration cycles. Characterization and calibration of these parameters inside the device would reduce the test time and cost considerably. In this paper, we propose a technique to measure the delay skew and the finite envelope bandwidth, two parameters that can be digitally calibrated, based on the measurement of the output of the receiver in the loop-back mode. Simulation and hardware measurement results show that the proposed technique can characterize the targeted impairments in the polar transmitter accurately.

**[378] A Flexible BIST Strategy for SDR Transmitters**

Emanuel Dogaru, Filippe Vincio dos Santos, William Rebeschini

Software-defined radio (SDR) development aims for increased speed and flexibility. The advent of these system-level requirements on the physical layer (PHY) access hardware is leading to more complex architectures, which together with higher levels of integration pose a challenging problem for product testing. For radio units that must be field-upgradeable without specialized equipment, built-in self-test (BIST) schemes are arguably the only way to ensure continued compliance to specifications. In this paper we introduce a lookback RF BIST technique that uses periodically nonuniform sampling (PNS2) of the transmitter (TX) output to evaluate compliance to spectral mask specifications. No significant hardware costs are incurred due to the re-use of available RX resources (I/Q ADCs, DSP, GPP, etc.). Simulation results of an homodyne TX demonstrate that adjacent channel power ratio (ACPR) can be accurately estimated. Future work will consist in validating our lookback RF BIST architecture on an in-house SDR testbed.

**Keywords - BIST, In-Field Test, Periodically Nonuniform Sampling, Software Radios, Mixed Signal RF Test, Subsampling, Spectral Mask Estimation**

---

**[379] Sigma-Delta Testability for Pipeline A/D Converters**

Antonio Gines, Gildas Leger

Pipeline Analog to Digital Converters (ADCs) are widely used in applications that require medium to high resolution at high acquisition speed. Despite of their quite simple working principles, they usually form rather complex mixed-signal blocks, particularly if digital correction and calibration are considered. As a result, pipeline converters are difficult to test and diagnose. In this paper we propose to reconfigure the internal multiplying DACs (MDACs) that perform residue amplifications as integrators, each one with an analog and a digital input. In this way, we can reuse consecutive pipeline stages to form ΔΣ modulators, with very reduced area overhead. We thus get an on-chip DC (low-frequency) probe with a digital 1-bit output that does not require any extra pin. In addition, digital test techniques developed for ΔΣ modulators may be used to enhance the diagnosing capabilities. An industrial 1.8V 15-bit 100Msps pipeline ADC that had previously been fully validated in a 0.18µm CMOS process is used as a case of study for the introduction of the ΔΣ modifications.

---

**12.8 Panel: Future SoC Verification Methodology: UVM Evolution or Revolution?**

Organizer: Alex Gonyaichev, IBM Research - Haifa, IL
Chair: Rolf Drechsler, University of Bremen/DFKI DE

**[380] Future SoC Verification Methodology: UVM Evolution or Revolution?**

Rolf Drechsler, Christophe Chevalia, Franco Fummi, Alan J. Hu, Ronny Morad, Frank Schirrmeister, Alex Gonyaichev

With increasing design complexity system on chip (SoC) verification is becoming a more and more important and challenging aspect of the overall development process. The universal verification methodology (UVM) is thereby a common solution to this problem; although it still keeps some problems unsolved. In this paper leading experts from industry (both users and vendors) and academia will discuss the future of SoC verification methodology.

Source URL: https://www.date-conference.com/conference/proceedings/2014/abstracts

Links:
