2.2: Validation of Modern Microprocessors

Moderators: D Grosse, Bremen U, DE; V Bertacco, U of Michigan, US

Automated Generation of Directed Tests for Transition Coverage in Cache Coherence Protocols [p. 3]

X Qin and P Mishra

Processes with multiple cores and complex cache coherence protocols are widely employed to improve the overall performance. It is a major challenge to verify the correctness of a cache coherence protocol since the number of reachable states grows exponentially with the number of cores. In this paper, we propose an efficient test generation technique, which can be used to achieve full state and transition coverage in simulation based verification for a wide variety of cache coherence protocols. Based on effective analysis of the state space structure, our method can generate more efficient test sequences (50% shorter) compared with tests generated by breadth first search. Moreover, our proposed approach can generate tests on-the-fly due to its space efficient design.
2.3: Memory System Optimization

Moderators: T Austin, EECS, U of Michigan, US; C Silvano, Politecnico di Milano, IT

Chip multiprocessing is key to Mobile and high-end Embedded Computing. It requires sophisticated multilevel hierarchies where private and shared caches coexist. It relies on hardware support to implicitly manage relaxed program order and write atomicity so as to provide well-defined shared-memory semantics (captured by the axioms of a memory consistency model) at the hardware-software interface. This paper addresses the problem of checking if an executable representation of the memory system complies with a specified consistency model. Conventional verification techniques encode the axioms as edges of a single directed graph, infer extra edges from memory traces, and indicate an error when a cycle is detected. Unlike them, we propose a novel technique that decomposes the verification problem into multiple instances of an extended bipartite graph matching problem. Since the decomposition was judiciously designed to induce independent instances, the target problem can be solved by a parallel verification algorithm. Our technique, which is proven to be complete for several memory consistency models, outperformed a conventional checker for a suite of 2400 randomly-generated use cases. On average, it found a higher percentage of faults (90%) as compared to that checker (69%) and did it, on average, 272 times faster.

Generating Instruction Streams Using Abstract CSP [p. 15]

Y Katz, M Rimon and A Ziv

One of the challenges that processor level stimuli generators are facing is the need to generate stimuli that exercise microarchitectural mechanisms deep inside the verified processor. These scenarios require specific relations between the instructions participating in them. We present a new approach for processor-level scenario generation. The approach is based on creating an abstract constraint satisfaction problem, which captures the essence of the requested scenario. The generation of stimuli is done by interleaving between progress in the solution of the abstract CSP and generation of instructions. Compared with existing solutions of scenario generation, this approach yields improved coverage and reduced generation fail rate.

A Cycle-Approximate, Mixed-ISA Simulator for the KAHRISMA Architecture [p. 21]

T Stripf, R Koenig and J Becker

Processor architectures that are capable to reconfigure their instruction set and instruction format dynamically at run time offer a new flexibility exploiting instruction level parallelism vs. thread level parallelism. Based on the characteristics of an application or thread the instruction set architecture (ISA) can be adapted to increase performance or reduce resource/power consumption. To benefit from this run-time flexibility automatic selection of an appropriate ISA for each function of a given application is envisioned. This demands a cycle-accurate simulator that is capable of measuring the performance characteristics of an ISA dependent on the target application. However, simulation speed of a cycle-accurate simulator of our reconfigurable VLIW-like processor instances featuring dynamic operation execution would become relatively slow due to the superscalar-like micro-architecture. Within this paper we address this problem by presenting our cycle-approximate simulator approach containing a heuristic dynamic operation execution and memory model that provides a good trade-off between performance and accuracy. Additionally, the simulator features measurement of instruction level parallelism (ILP) that could be theoretically exploited by VLIW processor instances running on our architecture. The theoretical ILP could be used as an indicator for the ISA selection process without the need to simulate any combination of the different ISAs and applications.

A Clustering-Based Scheme for Concurrent Trace in Debugging NoC-Based Multicore Systems [p. 27]

J Gao, J Wang, Y Han, L Zhang and X Li

Concurrent trace is an emerging challenge when debugging multicore systems. In concurrent trace, trace buffer becomes a bottleneck since all trace sources try to access it simultaneously. In addition, the on-chip interconnection fabric is extremely high hardware cost for the distributed trace signals. In this paper, we propose a clustering-based scheme which implements concurrent trace for debugging Network-on-Chip (NoC) based multicore systems. In the proposed scheme, a unified communication framework eliminates the requirement for interconnection fabric which is only used during debugging. With clustering scheme, multiple concurrent trace sources can access distributed trace buffer via NoC under bandwidth constraint. We evaluate the proposed scheme using Booksim and the results show the effectiveness of the proposed scheme.

CACTI-3DD: Architecture-level Modeling for 3D Die-stacked DRAM Main Memory [p. 33]

K Chen, S Li, N Muralimanohar, J H Ahn, J B.Brockman and N P Jouppi
Emerging 3D die-stacked DRAM technology is one of the most promising solutions for future memory architectures to satisfy the ever-increasing demands on performance, power, and cost. This paper introduces CACTI-3DD, the first architecture-level integrated power, area, and timing modeling framework for 3D die-stacked off-chip DRAM main memory. CACTI-3DD includes TSV models, improves models for 2D off-chip DRAM main memory over current versions of CACTI, and includes 3D integration models that enable the analysis of a full spectrum of 3D DRAM designs from coarse-grained rank-level 3D stacking to bank-level 3D stacking. CACTI-3DD enables an in-depth study of architecture-level tradeoffs of power, area, and timing for 3D die-stacked DRAM designs. We demonstrate the utility of CACTI-3DD in analyzing design trade-offs of emerging 3D die-stacked DRAM main memories. We find that a coarse-grained 3D DRAM design that stacks canonical DRAM dies can only achieve marginal benefits in power, area, and timing compared to the original 2D design. To fully leverage the huge internal bandwidth of TSVs, DRAM dies must be re-architected, and system implications must be considered when building 3D DRAMs with redesigned 2D planar DRAM dies. Our results show that the 3D DRAM with re-architected DRAM dies achieves significant improvements in power and timing compared to the coarse-grained 3D die-stacked DRAM.

Keywords: 3D architecture, DRAM, TSV, Main memory, Modeling

S Stipic, S Tomic, F Zyulkyarov, A Cristal, O Unsal and M Valero

In this paper we introduce TagTM, a Software Transactional Memory (STM) system augmented with a new hardware mechanism that we call GTags. GTags are new hardware cache coherent tags that are used for fast meta-data access. TagTM uses GTags to reduce the cost associated with accesses to the transactional data and corresponding metadata. For the evaluation of TagTM, we use the STAMP TM benchmark suite. In the average case TagTM provides a speedup of 7-15% (across all STAMP applications), and in the best case shows up to 52% speedup of committed transaction execution time (for SSCA2 application).

Y-T Chen, J Cong, H Huang, B Liu, C Liu, M Potkonjak and G Reinman

The recent development of non-volatile memory (NVM), such as spin-torque transfer magnetoresistive RAM (STT-RAM) and phase-change RAM (PRAM), with the advantage of low leakage and high density, provides an energy-efficient alternative to traditional SRAM in cache systems. We propose a novel reconfigurable hybrid cache architecture (RHC), in which NVM is incorporated in the last-level cache together with SRAM. RHC can be reconfigured by powering on/off SRAM/NVM arrays in a way-based manner. In this work, we discuss both the architecture and circuit design issues for RHC. Furthermore, we provide hardware-based mechanisms to dynamically reconfigure RHC on-the-fly based on the cache demand. Experimental results on a wide range of benchmarks show that the proposed RHC achieves an average 63%, 48% and 25% energy saving over non-reconfigurable SRAM-based cache, non-reconfigurable hybrid cache, and reconfigurable SRAM-based cache, while maintaining the system performance (at most 4% performance overhead).

M D Gomony, C Weis, B Akesson, N Wehn and K Goossens

The performance and power consumption of mobile DRAMs (LPDDRs) depend on the configuration of system-level parameters, such as operating frequency, interface width, request size, and memory map. In mobile systems running both real-time and non-real-time applications, the memory configuration must satisfy bandwidth requirements of real-time applications, meet the power consumption budget, and offer the best average-case execution time to the non-real-time applications. There is currently no well-defined methodology for selecting a suitable memory configuration for real-time mobile systems. The worst-case bandwidth, average-case execution time, and power consumption of mobile DRAMs across generations have furthermore not been investigated. This paper has two main contributions. 1) We analyze the worst-case bandwidth, average-case execution time, and power consumption of mobile DRAMS across three generations: LPDDR, LPDDR2 and Wide-IO-based 3D-stacked DRAM. 2) Based on our analysis, we propose a methodology for selecting memory configurations in real-time mobile systems. We show that LPDDR (32-bit IO), LPDDR2 (32-bit IO) and 3D-DRAM (128-bit IO) provide worst-case bandwidth up to 0.75 GB/s, 1.6 GB/s and 3.1 GB/s, respectively. We furthermore show for an H.263 decoder that LPDDR2 and 3D-DRAM reduce power consumption with up to 25% and 67%, respectively, compared to LPDDR, and reduce the execution time with up to 18% and 25%.

2.4: Architectures and Efficient Designs for Automotive and Energy-Management Systems

Moderators: C Sebeke, Bosch, DE; G Merrett, Southampton U, UK

[23] Using Timing Analysis for the Design of Future Switched Based Ethernet Automotive Networks [p. 57]
J Rox, R Ernst and P Giusto

In this paper, we focus on modeling and analyzing multi-cast and broadcast traffic latencies on switch-level within an Ethernet-based communication network for automotive applications. The analysis is performed adapting existing worst/best case schedulability analysis
2.5: Physical Design for Low-Power

Moderators: J Teich, Erlangen-Nurenberg U, DE; W Fomaciari, Politecnico di Milano, IT

Electromigration (EM) has been indicated as the killer effect for copper interconnects. ITRS projections show that for future technologies (22nm and beyond) the on-chip current demand will exceed the physical limit copper metal wires can tolerate. This represents a serious limitation for the design of power distribution networks of next generation ICs. New carbon nanomaterials, governed by ballistic transport, have shown higher immunity to EM, thereby representing potential candidate to replace copper. In this paper we make use of compact conductance models to benchmark Graphene Nanoribbons (GNRs) against copper. The two materials have been used to route a state-of-the-art multi-level power-grid architecture obtained through an industrial 45nm physical design flow. Although the adopted design style is optimized for metal grids, results obtained using our simulation framework show that GNRs, if properly sized, can outperform copper, thus allowing the design of reliable circuits with reduced IR-drop penalties.

[27] IR-Drop Analysis of Graphene-Based Power Distribution Networks [p. 81]
S Miryala, A Calimera, E Macii and M Poncino
2.6: Optimized Utilization of Embedded Platforms

Moderators: F Slomka, Ulm U, DE; O Bringmann, FZI Karlsruhe, DE

[31] Fast and Lightweight Support for Nested Parallelism on Cluster-Based Embedded Many-Cores [p. 105]
A Marongiu, P Burgio and L Benini

Several recent many-core accelerators have been architected as fabrics of tightly-coupled shared memory clusters. A hierarchical interconnection system is used - with a crossbar-like medium inside each cluster and a network-on-chip (NoC) at the global level - which make memory operations nonuniform (NUMA). Nested parallelism represents a powerful programming abstraction for these architectures, where a first level of parallelism can be used to distribute coarse-grained tasks to clusters, and additional levels of fine-grained parallelism can be distributed to processors within a cluster. This paper presents a lightweight and highly optimized support for nested parallelism on cluster-based embedded many-cores. We assess the costs to enable multi-level parallelization and demonstrate that our techniques allow to extract high degrees of parallelism.

[32] A Divide and Conquer Based Distributed Run-time Mapping Methodology for Many-Core Platforms [p. 111]
I Anagnostopoulos, A Bartzas, G Kathareios and D Soudris

Real-time applications are raising the challenge of unpredictability. This is an extremely difficult problem in the context of modern, dynamic, multiprocessor platforms which, while providing potentially high performance, make the task of timing prediction extremely difficult. In this
In the recent years, commodity solid-state disks have started adopting powerful controllers and implemented page-level mapping for flash management. However, many of these models still use primitive garbage-collection algorithms, because prior approaches do not scale up with the dramatic increase of flash capacity. This study introduces Dual Greedy for garbage collection in page-level mapping. Dual Greedy identifies page-accurate data hotness using only block-level information, and adaptively switches its preference of victim selection between block space utilization and block stability. It can run in constant time and use very limited RAM space. Our experimental results show that Dual Greedy outperforms existing approaches in terms of garbage-collection overhead, especially with large flash blocks.
application programs are host-compiled and use a single transaction to transfer a data block. A strategy is presented that efficiently estimates the timing of block transactions based on the profile library. It is the first method that takes into account caching effects that influence the timing of block transactions. Moreover, it ensures overall timing accuracy when integrated in other SW timing tools for full system simulation. Experimental results show that the block transactions are accurately timed, with average error less than 1%. At the same time, the simulation gain can be up to three orders of magnitude.

W Chen, X Han and R Doemer

At the Electronic System Level (ESL), design validation often relies on discrete event (DE) simulation. Recently, parallel simulators have been proposed which increase simulation speed by using multiple cores available on today's PCs. However, the total order of time in DE simulation is a bottleneck that severely limits the benefits of parallel simulation. This paper presents a new out-of-order simulator for multi-core parallel DE simulation of hardware/software designs at any abstraction level. By localizing the simulation time and carefully handling events at different times, a system model can be simulated following a partial order of time. Subject to automatic static data analysis at compile time and table-based decisions at run time, threads can be issued early which reduces the idle time of available cores. Our experiments show high performance gains in simulation speed with only a small increase of compile time.

H-Y Lin, C-Y Wang, S-C Chang, Y-C Chen, HM Chou, C-Y Huang, Y-C Yang and C-C Shen

Mutation Analysis (MA) is a fault-based simulation technique that is used to measure the quality of testbenches in error (mutant) detection. Although MA effectively reports the living mutants to designers, it suffers from the high simulation cost. This paper presents a probabilistic MA preprocessing technique, Error Propagation Analysis (EPA), to speed up the MA process. EPA can statically estimate the probability of the error propagation with respect to each mutant for guiding the observation-point insertion. The inserted observation-points will reveal a mutant's status earlier during the simulation such that some useless testcases can be discarded later. We use the mutant model from an industrial EDA tool, Certitude, to conduct our experiments on the OpenCores' RT-level designs. The experimental results show that the EPA approach can save about 14% CPU time while obtaining the same mutant status report as the traditional MA approach.

B Mammo, D Chatterjee, D Pidan, A Nahir, A Ziv, R Morad and V Bertacco

Simulation-based functional verification is the key validation methodology the industry. The performance of logic simulators, however, is not sufficient to attain acceptable verification coverage on large industrial designs within the time-frame available. Acceleration platforms are a valuable addition to the verification effort in that they can provide much higher coverage in less time. Unfortunately, these platforms do not provide the rich checking capability of software-based simulation. We propose a novel solution to deploy those complex checkers, typical of simulation-based environments, onto acceleration platforms. To this end, checkers must be transformed into synthesizable, compact logic blocks with bug-detection capabilities similar to that of their software counterparts. Our "approximate checkers" trade off logic complexity with bug detection accuracy by leveraging novel techniques to approximate complex software checkers into small synthesizable hardware blocks, which can be simulated along with the design on an acceleration platform. We present a general checker taxonomy, propose a range of approximation techniques based on a checker's characteristic and provide metrics for evaluating its bug detection capabilities.

3.3: Industrial Design Methodologies

Moderators: A Jerraya, CEA, FR; R Zafalon, STMicroelectronics, IT

D Steinbach

Cassidian® is working on modeling guidelines. We present our approach and report first results and findings to illustrate progress and direction of our work.

Systems Engineering; Model Based; Guidelines; Rules

N Battezzati, S Colazzo, M Maffione and L Senepa

Today many industrial applications require object recognition and tracking capabilities. Feature-based algorithms are well-suited for such operations and, among all, Speeded Up Robust Features (SURF) algorithm has been proved to achieve optimal results. However, when high-
Precision and real-time requirements come together, a dedicated hardware is necessary to meet them. In this paper we present a novel architecture for implementing SURF algorithm in FPGA, along with experimental results for different industrial applications.

Moderators: G Palermo, Politecnico di Milano, IT; M Poncino, Politecnico di Torino, IT

3.4: Large-Scale Energy and Thermal Management

We present in this paper NOCEVE an industrial Network on Chip (NoC) emulation and verification environment on industrial large scale multi-FPGA emulation platform for billion cycle application. It helps designer to improve system performance by the analysis of traffic distribution and balance through the network on chip. The hardware monitoring network is generated by another commercial NoC design tool. It consists of traffic collectors, which is reconfigurable to collect different traffic information such as packet latency and throughput. The traffic information is collected during real application execution on FPGA platform and it is sent through monitoring network on FPGA and then PCI bright board back to host computer for real-time visualization or post-execution data analysis. NOCEVE is the first industrial NoC emulation and verification environment for billion cycle applications.

Keywords: emulation, FPGA, NoC, verification.

Investigating the Effects of Inverted Temperature Dependence (ITD) on Clock Distribution Networks [p. 165]
A Sassone, A Calimera, A Macii, E Macii, M Poncino, R Goldman, V Melikyan, E Babayan and S Rinaudo

The aggressive scaling of CMOS technology toward nanometer lengths contributed to the surfacing of many effects that were not appreciable at the micrometer regime. Among them, Inverted Temperature Dependence (ITD) is certainly the most unusual. It manifests itself as a speed up of CMOS gates when the temperature increases, resulting in a reversal of the worst-case condition, i.e., CMOS gates show the largest delay at low temperatures. On the other hand, for metal interconnects an high temperature still holds as worst case condition. The two contrasting behaviors may invalidate the results obtained through standard design flow which do not consider temperature as an explicit variable in their optimizations. In this paper we focus on the impact of ITD on clock distribution networks (CDN), whose function is vital to guarantee the synchronization among physically spaced sequential components of digital circuits. Using our simulation framework, we characterized the thermal behavior of a clock tree mapped onto an industrial 65nm CMOS technology and obtained using a standard synthesis tool. Results demonstrate the presence of ITD at low operating voltages and open new potential research scenarios into the EDA field.

Challenges in Verifying an Integrated 3D Design [p. 167]
T G Yip, C Y Hung and V Iyengar

The integrated 3D configuration considered in this study includes a silicon die on one side of an organic interposer and a different die on the other side. The three parts are from three different design environments, each has its own database and description language not compatible with the other two. The incompatibility triggered a search for a new methodology for the physical verification of the 3D configuration. Application scripts were developed and successfully used to verify the physical connections within the complex design of the interposer, which accommodates 1600 signals and 12,000 traces for connecting the signals between the two chips. The layout of 56,000 vias for power and signal was also verified to meet the requirements for the manufacturing of the organic interposer.

Keywords: 3D design; verification; integration; system design.

Multiple-Source and Multiple-Destination Charge Migration in Hybrid Electrical Energy Storage Systems [p. 169]
Y Wang, Q Xie, M Pedram, Y Kim, N Chang and M Poncino

Hybrid electrical energy storage (HEES) systems consist of multiple banks of heterogeneous electrical energy storage (EES) elements that are connected to each other through the Charge Transfer Interconnect. A HEES system is capable of providing an electrical energy storage means with very high performance by taking advantage of the strengths (while hiding the weaknesses) of individual EES elements used in the system. Charge migration is an operation by which electrical energy is transferred from a group of source EES elements to a group of destination EES elements. It is a necessary process to improve the HEES system’s storage efficiency and its responsiveness to load demand changes. This paper is the first to formally describe a more general charge migration problem, involving multiple sources and multiple destinations. The multiple-source, multiple-destination charge migration optimization problem is formulated as a nonlinear programming (NLP) problem where the goal is to deliver a fixed amount of energy to the destination banks while maximizing the overall charge migration efficiency and not depleting the available energy resource of the source banks by more than a given percentage. The constraints for the optimization problem are the energy conservation relation and charging current constraints to ensure that charge migration will meet a given deadline. The formulation correctly accounts for the efficiency of chargers, the rate capacity effect of batteries, self-discharge currents and internal resistances of EES elements, as well as the terminal voltage variation of EES elements as a function of their state of charges (SoC’s).
An efficient algorithm to find a near-optimal migration control policy by effectively solving the above NLP optimization problem as a series of quasiconvex programming problems is presented. Experimental results show significant gain in migration efficiency up to 35%.

Keywords: hybrid electrical energy storage system; charge management; charge migration

Benefits of Green Energy and Proportionality in High Speed Wide Area Networks Connecting Data Centers [p. 175]

B Aksanli, T S Rosing and I Monga

Many companies deploy multiple data centers across the globe to satisfy the dramatically increased computational demand. Wide area connectivity between such geographically distributed data centers has an important role to ensure both the quality of service, and, as bandwidths increase to 100Gbps and beyond, as an efficient way to dynamically distribute the computation. The energy cost of data transmission is dominated by the router power consumption, which is unfortunately not energy proportional. In this paper we not only quantify the performance benefits of leveraging the network to run more jobs, but also analyze its energy impact. We compare the benefits of redesigning routers to be more energy efficient to those obtained by leveraging locally available green energy as a complement to the brown energy supply. Furthermore, we design novel green energy aware routing policies for wide area traffic and compare to state-of-the-art shortest path routing algorithm. Our results indicate that using energy proportional routers powered in part by green energy along with our new routing algorithm results in 10x improvement in per router energy efficiency with 36% average increase in the number of jobs completed.

Keywords: green energy, network, energy proportional, routing.


A Bartolini, M Sadri, J-N Furst, A K Coskun and L Benini

Dynamic frequency and voltage scaling (DVFS) techniques have been widely used for meeting energy constraints. Single-chip many-core systems bring new challenges owing to the large number of operating points and the shift to message passing interface (MPI) from shared memory communication. DVFS, however, has been mostly studied on single-chip systems with one or few cores, without considering the impact of the communication among cores. This paper evaluates the impact of frequency scaling on the performance and power of many-core systems with MPI. We conduct experiments on the Single-Chip Cloud Computer (SCC), an experimental many-core processor developed by Intel. The paper first introduces the run-time monitoring infrastructure and the application suite we have designed for an in-depth evaluation of the SCC. We provide an extensive analysis quantifying the effects of frequency perturbations on performance and energy efficiency. Experimental results show that runtime communication patterns lead to significant differences in power/performance tradeoffs in many-core systems with MPI.


G Liu, M Fan and G Quan

With the high integration density and complexity of the modern multi-core platform, thermal problems become more and more significant for both the manufacture and system designer. Dynamic thermal management technique is one effective and efficient way to mitigate and avoid thermal emergences. In this paper, we propose a novel predictive dynamic thermal management algorithm to maximize the multi-core system throughput while satisfying the peak temperature constraints. Different from the conventional approaches, we found that it is not necessarily always a good choice to migrate a hot task to the core with the lowest temperature. Instead, in our algorithm, we develop a new temperature prediction technique and migration scheme that take the local temperature of a core as well as the impacts from neighboring cores into considerations. According to our experiment results on a practical Intel desktop platform, the proposed algorithm can significantly improve the throughput compared with the conventional approach.

3.5: PANEL - Key Challenges for Next Generation Computing

Moderator: R Riemenschneider, European Commission, BE

The exponential growth in IT made possible through Moore's law for several decades has been surpassed by the demand for computing. Future high performance computing (HPC) systems are considered an area also relevant for traditional safety-critical embedded systems like automotive and aerospace. HPC could also benefit from experiences in embedded computing in terms of fault-tolerant run-time environment (RTE) architectures with high degree of reliability and dependability. The panel objective is to explore interdisciplinary technologies cutting across multi-core computing systems, dependable computing and high performance computing. The panel brings together industry and academia from so far fragmented domains such as real-time embedded system engineering and HPC architectures.

3.6: Model-Based Design and Verification for Embedded Systems

Moderators: W Yi, Uppsala U, SE; S Ben Salem, Verimag Laboratory, FR
Improving Reliability and Yield in Advanced Technologies

Moderators: S Nassif, IBM, US; S Khursheed, Southampton U, UK

3.7: Improving Reliability and Yield in Advanced Technologies

The two-player mean-payoff game is a well-known game theoretic model that is widely used, for instance in economics and control theory. For controller synthesis, a controller is modeled as a player while the environment, or plant, is modeled as the opponent player (adversary). Synthesizing an optimal controller that satisfies a given criterion corresponds to finding a winning strategy for the controller player. Emerging streaming applications (audio, video, communication, etc.) for embedded systems exhibit both input sensitive and controller sensitive runtime behavior, where the controller's role is runtime management or scheduling. Embedded controllers need to be optimized for dynamic inputs, while guaranteeing throughput constraints. In this paper, we consider this design task for scenario- and resource-aware dataflow graphs that model streaming applications. Scenarios in these models capture classes of dynamic environment behavior. We demonstrate how to model and solve the controller synthesis problem by constructing a winning strategy in a two-player mean payoff throughput game.

Index Terms - Synchronous DataFlow, Maxplus Algebra, Game Theory, Policy Iteration

Verifying Timing Synchronization Constraints in Distributed Embedded Architectures

A C Rajeev, S Mohalik and S Ramesh

Correct functioning of automotive embedded controllers requires hard real-time constraints on a number of system parameters. To avoid costly design iterations, these timing constraints should be verified during the design stage itself. In this paper, we describe a formal verification technique for a class of timing constraints called timing synchronization constraints in the recent adaptation of AUTOSAR standard (WPII-1.2 Timing Subgroup, Release 4.0). These constraints require, unlike the well studied end-to-end latency constraint, simultaneous analysis of multiple task/message chains or multiple data items traversing through a task/message chain. We show that they can be analyzed by model-checking with finite-state monitors. We also demonstrate this method on a case-study from the automotive domain.

Task Implementation of Synchronous Finite State Machines

M Di Natale and H Zeng

Model-based design of embedded control systems using Synchronous Reactive (SR) models is among the best practices for software development in the automotive and aeronautics industry. SR models allow to formally verify the correctness of the design and to automatically generate the implementation code. This improves productivity and, more importantly, can ensure a correct software implementation (preserving the model semantics). Previous research focuses on the concurrent implementation of the dataflow part of SR models, including the optimization of the block-to-task mapping and communication buffer sizing. When the system also consists of blocks implementing finite state machines, as in modern modeling tools like Simulink and SCADE, the task implementation can be further optimized with respect to time and memory. In this paper we analyze problems and opportunities in the implementation of finite state machine subsystems. We define the constraints and efficient policies for the task implementation of such systems.

Enabling Dynamic Assertion-based Verification of Embedded Software through Model-driven Design

G Di Guglielmo, L Di Guglielmo, F Fummi and G Pravadelli

Assertion-based verification (ABV) is more and more used for verification of embedded systems concerning both HW and SW parts. However, ABV methodologies and tools do not apply to HW and SW components in the same way: for HW components, both static ABV and dynamic ABV are widely used; on the contrary, SW components are traditionally verified by means of static ABV, because dynamic approaches are based on simulation assumptions which could not be true during execution of general embedded SW and which cannot be controlled by the assertion language. This paper proposes to exploit model-driven design for guaranteeing such simulation assumptions. Then, it describes an ABV framework for embedded SW, that automatically synthesizes assertion checkers to verify the embedded SW accordingly to the simulation assumptions.

3.7: Improving Reliability and Yield in Advanced Technologies

Negative Bias Temperature Instability (NBTI) is a major source of transistor aging in scaled CMOS, resulting in slower devices and shorter lifetime. NBTI is strongly dependent on the input vector. Moreover, a considerable fraction of execution time of an application is spent to execute NOP (No Operation) instructions. Based on these observations, we present a novel NOP assignment to minimize NBTI effect, i.e. maximum NBTI relaxation, on the processors. Our analysis shows that NBTI degradation is more impacted by the source operands rather than...
3.8: HOT TOPIC - Design Automation Tools for Engineering Biological Systems

Moderator: J Madsen, DTU, DK

J Pontes, N Calazans and P Vivet

Similar to local variations and signal integrity problems, Single Event Effects (SEEs) are a new design concern for digital system design that arises in deep sub-micron technologies. In order to design reliable digital systems in such technologies, it is mandatory to precisely model and take into account SEEs. This paper proposes a new accurate design flow to model non-permanent SEE effects that can be applied at system level for reliable digital circuit design. Starting from low level SPICE-accurate simulations, SEEs are characterized, modeled and simulated in the digital design using commercial and well accepted standards and tools. The proposed design flow has been fully validated through a complete digital design, a cryptographic core implemented in a 32nm CMOS technology. Finally, using the SEE design flow, the paper presents some reliability impact analysis, both at standard cell level and design level.

Keywords-component: Single event effects, soft errors, radiation hardening.

Cross Entropy Minimization for Efficient Estimation of SRAM Failure Rate [p. 230]
M A Shahid

As the semiconductor technology scales down to 45nm and below, process variations have a profound effect on SRAM cells and an urgent need is to develop fast statistical tools which can accurately estimate the extremely small failure probability of SRAM cells. In this paper, we adopt the Importance Sampling (IS) based information theory inspired Minimum Cross Entropy method, to propose a general technique to quickly evaluate the failure probability of SRAM cells. In particular, we first mathematically formulate the failure of SRAM cells such that the concept of “Cross Entropy Distance” can be leveraged, and the distance between the ideal distribution for IS and the practical distribution for IS (which is used for generating samples), is well-defined. This cross entropy distance is now minimized resulting in a simple analytical solution to obtain the optimal practical distribution for IS, thereby expediting the convergence of estimation. The experimental results of a commercial 45nm SRAM cell demonstrate that for the same accuracy, the proposed method yields computational savings on the order of 17~50X over the existing state-of-the-art techniques.

Experimentally Driven Verification of Synthetic Biological Circuits [p. 236]
B Yordanov, E Appleton, R Ganguly, E A Gol, S B Can, S Bhatia, T Haddock, C Belta, D Densmore

We present a framework that allows us to construct and formally analyze the behavior of synthetic gene circuits from specifications in a high level language used in describing electronic circuits. Our back-end synthesis tool automatically generates genetic-regulatory network (GRN) topology realizing the specifications with assigned biological “parts” from a database. We describe experimental procedures to acquire characterization data for the assigned parts and construct mathematical models capturing all possible behaviors of the generated GRN. We delineate algorithms to create finite abstractions of these models, and novel analysis techniques inspired from model-checking to verify behavioral specifications using Linear Temporal Logic (LTL) formulae.

Genetic/Bio Design Automation for (Re-)Engineering Biological Systems [p. 242]
S Hassoun

Constructing biological circuits in a bottom-up modular fashion using design methodologies similar to those used in electronics has gained tremendous attention in the past decade. The end goal, however, is engineering biological systems and not only individual components in the context of pursuing applications useful in improving human health or enhancing the environment. This article reviews the basics of biological system design rooted in Metabolic Engineering and Systems Biology and outlines current system-level modeling, analysis, optimization, and synthesis with emphasis on some current bottlenecks in establishing more rigorous design tools and methodologies for engineering biological systems.

IP1: Interactive Presentations

Fast Cycle Estimation Methodology for Instruction-Level Emulator [p. 248]
In this paper, we propose a cycle estimation methodology for fast instruction-level CPU emulators. This methodology suggests achieving accurate software performance estimation at high emulation speed by utilizing a two-phase pipeline scheduling process: a static pipeline scheduling phase performed off-line before runtime, followed by an accuracy refinement phase performed at runtime. The first phase delivers a pre-estimated CPU cycle count while limiting impact on the emulation speed. The second phase refines the pre-estimated cycle count to provide further accuracy. We implemented this methodology on QEMU and compared cycle counts with a physical ARM CPU. Our results show the efficiency of the tradeoffs between emulation speed and cycle accuracy: cycle simulation error averages 10% while the emulation latency is 3.37 times that of original QEMU.

Verification of embedded multicore applications is crucial as these applications are deployed in many safety critical systems. Verification task is complicated by concurrency inherent in such applications. We use mutation testing to obtain a quantitative verification coverage metric for multicore applications developed using the new Multicore Communication API (MCAPI) standard. MCAPI is a lightweight API that targets heterogeneous multicore embedded systems. We developed a mutation coverage tool and performed several experiments on MCAPI applications. Our experiments show that mutation coverage is useful in measuring and improving the quality of the test suites and ultimately the quality of the multicore application.

Multithreaded processors increase throughput by executing multiple independent programs on a single pipeline. Simultaneous Multithreaded (SMT) processors execute multiple threads simultaneously thus add a significant dimension to the design complexity. Dealing with this complexity calls for extended and innovative design verification efforts. This paper develops an analytic model based SMT random test generation technique. SMT analytic model parameters are applied to create random tests with high utilization and increased contention. To demonstrate the methodology, parameters extracted from the PPC ISA and sample processor configurations are simulated on the SMT analytic model. The methodology focuses on exploiting data/control and structural hazards to guide the random test generator to create effective SMT tests.

Phase-change RAM (PCM) has the advantages of better scaling and non-volatility compared with the DRAM which is expected to face its scaling limit in the near future. There have been many studies on applying the PCM to main memory in order to complement or replace the DRAM. One common limitation of these studies is that they are based on synthetic PCM models. In our study, we investigate the feasibility and issues of applying a real PCM to main memory. In this paper, we report our case study of characterizing the PCM and evaluating its usefulness in the main memory. Our results show that the PCM/DRAM hybrid main memory with a modest DRAM size can give comparable performance to that of the DRAM only main memory. However, the hybrid memory with small DRAMs or large footprint programs can suffer from performance degradation due to the long latency of both PCM writes and write preemption penalty, which requires architectural innovations for exploiting the full potential of PCM write performance.
Camera-based driver assistance systems have attracted the attention of all major automotive manufacturers in the past several years and are increasingly utilized to differentiate a vendor's vehicles from its competitors. The calculation of depth information and Motion Estimation can be considered as two fundamental image processing applications in these systems, which have already been evaluated in diverse research scenarios. However, in order to push these computation-intensive features towards series integration, future in-vehicle implementations must adhere to the automotive industry's strict power consumption and cost constraints. As an answer to this challenge, this paper presents a high-performance FPGA-based dense block matching solution, which enables the calculation of both object motion and the extraction of depth information on shared hardware resources. This novel single-design approach significantly reduces the amount of logic resources required, resulting in valuable cost and power savings. The acquired sensor information can be fused into 3D positions with an associated 3D motion vector, which enables a robust perception of the vehicle’s environment. The modular implementation offers enhanced configuration features at design and execution time and achieves up to 418 GOPS at a moderate energy consumption of 10 Watts, providing a flexible solution for a future series integration.

Instrumented Medical Shoes (MSs) are equipped with a variety of sensors for measurement of quantities such as pressure, acceleration, and temperature which are often greatly beneficial in numerous diagnosis, monitoring, rehabilitation, and other medical tasks. One of primary limiting factors of MSs is their energy sensitivity. In order to overcome this limitation, we have developed an optimization intensive approach for energy harvesting. Our goal is to size and position a single piezoelectric transducer for energy generation in a medical shoe in such a way that maximal energy is collected and/or specified maximal voltage is achieved while collecting energy. We propose a scenario approach that provides statistically sound solution and evaluate our approach using our medical shoe simulator for subject specific energy harvesting and generic MS scavenging. We could get 3.7X energy gain compare to smallest size sensor and 1.3X energy gain compared to sensor with the size of a shoe.

FlexRay is likely to become the de-facto standard for upcoming in-vehicle communication. Efficient scheduling of the static and dynamic segment of the communication cycle in combination with the determination of more than 60 parameters that are part of the FlexRay protocol is a challenging task. This paper provides a formal analysis for interdependencies between the parameters as well as a scheduling approach for the static and dynamic segment. Experimental results give evidence of a significant interdependency between the subtasks such that a holistic scheduling approach becomes mandatory to provide high-quality FlexRay schedules. As a solution, this work introduces a complete functional FlexRay scheduling approach that takes parameter selection, allocation of messages to the static and dynamic segment, and concurrent scheduling into account. A real-world case study from the automotive domain gives evidence of efficiency and applicability of the proposed approach.

Efficiently managing the parallel execution of various application tasks onto a heterogeneous multi-core system consisting of a combination of processors and accelerators is a difficult task due to the complex system architecture. The management of reconfigurable multi-core systems which exploit dynamic and partial reconfiguration in order to, e.g. increase the number of processing elements to fulfill the performance demands of the application, is even more complicated. This paper presents a special virtualization layer consisting of one central server and several distributed computing clients to virtualize the complex and adaptive heterogeneous multi-core architecture and to autonomously manage the distribution of the parallel computation tasks onto the different processing elements.

Power consumption variability of both on-chip SRAMs and off-chip DRAMs is expected to continue to increase over the next decades. We opportunistically exploit this variability through a novel Variability-aware Memory Virtualization (VaMV) layer that allows programmers to partition their application's address space (through annotations) into virtual address regions and create mapping policies for each region. Each policy has different requirements (e.g., power, fault-tolerance) and is exploited by our dynamic memory management module (VaMVisor), which adapts to the underlying hardware, prioritizes the memory resources according to their characteristics (e.g., power

---

Optimization Intensive Energy Harvesting [p. 272]

M. Rofouei, M. A. Ghodrat, M. Potkonjak and A. Martinez Nova

Instrumented Medical Shoes (MSs) are equipped with a variety of sensors for measurement of quantities such as pressure, acceleration, and temperature which are often greatly beneficial in numerous diagnosis, monitoring, rehabilitation, and other medical tasks. One of primary limiting factors of MSs is their energy sensitivity. In order to overcome this limitation, we have developed an optimization intensive approach for energy harvesting. Our goal is to size and position a single piezoelectric transducer for energy generation in a medical shoe in such a way that maximal energy is collected and/or specified maximal voltage is achieved while collecting energy. We propose a scenario approach that provides statistically sound solution and evaluate our approach using our medical shoe simulator for subject specific energy harvesting and generic MS scavenging. We could get 3.7X energy gain compare to smallest size sensor and 1.3X energy gain compared to sensor with the size of a shoe.

Keywords: energy harvesting; medical shoes;

Designing FlexRay-based Automotive Architectures: A Holistic OEM Approach [p. 276]

P. Milbredt, M. Glass, M. Lukasiewycz, A. Steininger and J. Teich

FlexRay is likely to become the de-facto standard for upcoming in-vehicle communication. Efficient scheduling of the static and dynamic segment of the communication cycle in combination with the determination of more than 60 parameters that are part of the FlexRay protocol is a challenging task. This paper provides a formal analysis for interdependencies between the parameters as well as a scheduling approach for the static and dynamic segment. Experimental results give evidence of a significant interdependency between the subtasks such that a holistic scheduling approach becomes mandatory to provide high-quality FlexRay schedules. As a solution, this work introduces a complete functional FlexRay scheduling approach that takes parameter selection, allocation of messages to the static and dynamic segment, and concurrent scheduling into account. A real-world case study from the automotive domain gives evidence of efficiency and applicability of the proposed approach.

Virtualized On-Chip Distributed Computing for Heterogeneous Reconfigurable Multi-Core Systems [p. 280]

S. Werner, O. Dey, D. Goehringer, M. Huebner and J. Becker

Efficiently managing the parallel execution of various application tasks onto a heterogeneous multi-core system consisting of a combination of processors and accelerators is a difficult task due to the complex system architecture. The management of reconfigurable multi-core systems which exploit dynamic and partial reconfiguration in order to, e.g. increase the number of processing elements to fulfill the performance demands of the application, is even more complicated. This paper presents a special virtualization layer consisting of one central server and several distributed computing clients to virtualize the complex and adaptive heterogeneous multi-core architecture and to autonomously manage the distribution of the parallel computation tasks onto the different processing elements.

Keywords: multiprocessor, virtualization, parallel computing, FPGA, reconfigurable computing

VaMV: Variability-aware Memory Virtualization [p. 284]

L. A. D. Bathen, N. D. Dutt, A. Nicolau and P. Gupta

Power consumption variability of both on-chip SRAMs and off-chip DRAMs is expected to continue to increase over the next decades. We opportunistically exploit this variability through a novel Variability-aware Memory Virtualization (VaMV) layer that allows programmers to partition their application's address space (through annotations) into virtual address regions and create mapping policies for each region. Each policy has different requirements (e.g., power, fault-tolerance) and is exploited by our dynamic memory management module (VaMVisor), which adapts to the underlying hardware, prioritizes the memory resources according to their characteristics (e.g., power
consumption), and selectively maps data to the best-fitting memory resource (e.g., high-utilization data to low-power memory space). Our experimental results on embedded benchmarks show that VaMV is capable of reducing dynamic power consumption by 63% on average while reducing total execution time by an average of 34% by exploiting: 1) SRAM voltage scaling, 2) DRAM power variability, and 3) Efficient dynamic policy-driven variability-aware memory allocation.

Hybrid Simulation for Extensible Processor Cores [p. 288]
J Jovic, S Yakoushkin, L Murillo, J Eusse, R Leupers and G Ascheid

Due to their good flexibility-performance trade-off, Application Specific Instruction-set Processors (ASIPs) have been identified as a valuable component in modern embedded systems, especially the extensible ones, achieving good cost-efficiency trade-offs. Since the generation of the described hardware is usually automated to a high extent, in order to deliver an ASIP-based design in due time, developers are limited by the performance of the underlying simulation techniques for software development. On the other hand, the Hybrid Processor simulation technology (HySim), which enables dynamic run-time switching between native and instruction-accurate simulation, has reported high speed-up values for some fixed architectures. This paper presents enhanced HySim technology for extensible cores, based on a layered simulation infrastructure. This technology has shown a speed-up on a per-function basis of two orders of magnitude for a realistic MIMO OFDM benchmark on a multi-core platform with customized Xtensa cores by Tensilica.

Leveraging Reconfigurability to Raise Productivity in FPGA Functional Debug [p. 292]
Z Poulos, Y-S Yang, J Anderson, A Veneris and B Le

We propose new hardware and software techniques for FPGA functional debug that leverage the inherent reconfigurability of the FPGA fabric to reduce functional debugging time. The functionality of an FPGA circuit is represented by a programming bitstream that specifies the configuration of the FPGA’s internal logic and routing. The proposed methodology allows different sets of design internal signals to be traced solely by changes to the programming bitstream followed by device reconfiguration and hardware execution. Evidently, the advantage of this new methodology vs. existing debug techniques is that it operates without the need of iterative executions of the computationally-intensive design re-synthesis, placement and routing tools. In essence, with a single execution of the synthesis flow, the new approach permits a large number of internal signals to be traced for an arbitrary number of clock cycles using a limited number of external pins. Experimental results using commercial FPGA vendor tools demonstrate productivity (i.e. run-time) improvements of up to 30x vs. a conventional approach to FPGA functional debugging. These results demonstrate the practicality and effectiveness of the proposed approach.

MOUSSE: Scaling MOdelling and Verification to Complex HeterogeneoUS Embedded Systems Evolution [p. 296]
M Becker, G B G Defo, F Fummi, W Mueller, G Pravadelli and S Vinco

This work proposes an advanced methodology based on an open source virtual prototyping framework for verification of complex Heterogeneous Embedded Systems (HES). It supports early rapid modelling of complex HES through smooth refinements, an open interface based on IP-XACT extensions for secure composition of HES components, and automatic testbench generation over different abstraction levels.

Runtime Power Gating in Caches of GPUs for Leakage Energy Savings [p. 300]
Y Wang, S Roy and N Ranganathan

In this paper, we propose a novel microarchitectural technique for runtime power gating caches of GPUs to save leakage energy. The L1 cache (private to a core) can be put at a low-leakage sleep mode when there are no ready threads to be scheduled and the L2 can be put into sleep mode when there is no memory request. The sleep mode is state-retentive that precludes the necessity to flush the caches after they are woken up. The primary reason for the effectiveness of this technique lies in the fact that the latency of detecting cache inactivity, putting a cache to sleep and waking it up before it is accessed is completely hidden microarchitecturally. The technique incurs insignificant overheads in terms of power and area. Experiments were performed using the GPGPU-Sim simulator on benchmarks that was set up using the CUDA framework. The power and latency modeling of the cache arrays for measuring the wake-up latency and the breakeven periods is performed using 32-nm SOI IBM technology model. Based on experiments on 16 different GPU workloads, the average energy savings achieved by proposed technique is 54%.

Keywords - power gating, GPU, cache, SRAM, leakage power

Automatic Generation of Functional Models for Embedded Processor Extensions [p. 304]
F Sun

Early architectural exploration and design validation are becoming increasingly important for multi-processor systems-on-chip (MPSoC) designs. Native functional simulations can provide orders of magnitude in speedup over cycle or instruction level simulations but often require
Routing Solutions for Upcoming NoC Challenges

Moderators: J Flich, UP Valencia, ES; M Palesi, Kore U, IT

4.2: Routing Solutions for Upcoming NoC Challenges

In this work, we present a tool called NATIVESIM to automatically generate the functional models for embedded processor extensions. We provide a mechanism to address the challenge of modeling a subset of the processor architecture, with no visibility to the rest of the processor. We illustrate the problem of modeling the processor extensions when the endianness of the target processor is different from the host system and provide a solution to it. Experiments on several benchmark programs indicate that native execution of the target application with the functional models of the processor extensions can achieve large simulation run-time speedup over simulations based on either cycle accurate models (up to 14102x with an average of 3924x) or compiled functional models of an entire processor (up to 103x with an average of 31.6x).

Simulink/Stateflow (SL/SF) is the primary modeling notation for the development of control systems in automotive and aerospace industries. In model based testing, test cases derived from a design model are used to show model-code conformance. Safety standards such as ISO 26262 recommend model based testing to show the conformance of a software with the corresponding model. From our experiments with various test generation techniques, we have observed that their coverage capabilities are complementary in nature. With this observation in mind, we have developed a new tool called SmartTestGen which integrates different test generation techniques. In this paper, we discuss SmartTestGen and the different test generation techniques utilized - random testing, constraint solving, model checking and heuristics. We experimented with 20 production-quality SL/SF models and compared the performance of our tool with that of two prominent commercial tools.

Facing a growing complexity, embedded systems design relies on model-based approaches to ease the exploration of a design space. A key aspect of such exploration is performance evaluation, mainly depending on usage of the hardware resources. In model-driven engineering, hardware resources usage is often approximated by static properties. In this paper, we propose an extensible modeling framework, to describe with different levels of detail the hardware resource usage. Our method relies on the AADL to describe the whole system, and SystemC to refine the execution platform description. In this paper we expose how we generate and compose SystemC models from the execution platform model described in AADL. We also present promising experimental results obtained on an avionics use-case.

In this paper, we present RAG, an efficient Reliability Analysis tool based on Graphics processing units (GPU). RAG is a fault injection based parallel stochastic simulator implemented on a state-of-the-art GPU. A two-stage simulation framework is proposed to exploit the high computation efficiency of GPUs. Experimental results demonstrate the accuracy and performance of RAG. An average speedup of 412x and 198x is achieved compared to two state-of-the-art CPU-based approaches for reliability analysis.

Congestion occurs frequently in Networks-on-Chip when the packets demands exceed the capacity of network resources. Congestion-aware routing algorithms can greatly improve the network performance by balancing the traffic load in adaptive routing. Commonly, these algorithms either rely on purely local congestion information or take into account the congestion conditions of several nodes even though their statuses might be out-dated for the source node, because of dynamically changing congestion conditions. In this paper, we propose a method to utilize both local and non-local network information to determine the optimal path to forward a packet. The non-local information is gathered from the nodes that not only are more likely to be chosen as intermediate nodes in the routing path but also provide up-to-date information to a given node. Moreover, to collect and deliver the non-local information, a distributed propagation system is presented.

An MILP-Based Aging-Aware Routing Algorithm for NoCs

K Bhardwaj, K Chakraborty and S Roy
Network-on-Chip (NoC) architectures have emerged as a better replacement of the traditional bus-based communication in the many-core era. However, continuous technology scaling has made aging mechanisms such as Negative Bias Temperature Instability (NBTI) and electromigration primary concerns in NoC design. In this paper, we propose a novel system-level aging model to model the effects of asymmetric aging in NoCs. We observe a critical need of a holistic aging analysis, which, when combined with power-performance optimization, poses a multi-objective design challenge. To solve this problem, we propose a Mixed Integer Linear Programming (MILP)-based aging-aware routing algorithm that optimizes the various design constraints using a multi-objective formulation. After an extensive experimental analysis using real workloads, we observe a 62.7%, 46% average overhead reduction in network latency and Energy-Delay-Product-Per-Flit (EDPPF) and a 41% improvement in Instructions Per Cycle (IPC) using our aging-aware routing algorithm.

AFRA: A Low Cost High Performance Reliable Routing for 3D Mesh NoCs [p. 332]
S Akbari, A Shafiee, M Fathy and R Berangi

Three-dimensional network-on-chips are suitable communication fabrics for high-density 3D many-core ICs. Such networks have shorter communication hop count, compared to 2D NoCs, and enjoy fast and power efficient TSV wires in vertical links. Unfortunately, the fabrication process of TSV connections has not matured yet, which results in poor vertical links yield. In this work, we address this challenge and introduce AFRA, a deadlock-free routing algorithm for 3D mesh-based NoCs that tolerates faults on vertical links. AFRA is designed to be simple, high performance, and robust. The simplicity is achieved by applying ZXY and XZXY routings in the absence and presence of fault, respectively. Furthermore, AFRA, as will be proved, is deadlock-free when all vertical faulty links have the same direction. This enables the routing to save virtual channels for performance rather than sacrificing them for deadlock avoidance. Finally, AFRA provides robustness, which means supporting connection for all possible pairs of communicating nodes in high fault rates. AFRA is evaluated, though cycle accurate network simulation, and is compared with planar adaptive routing. Results reveal that AFRA significantly outperforms planar adaptive routing in both synthetic and real traffic patterns. In addition, the robustness of AFRA is calculated analytically.

4.3: Industrial Embedded System Design
Moderators: F Clermidy, CEA-Leti, FR; T Simunic Rosing, UC San Diego, US

E Patti, A Acquaviva, F Abate, A Osello, A Cucuccio, M Jahn, M Jentsch and E Macii

One of the major challenges in today's economy concerns the reduction in energy usage and CO₂ footprint in existing Public buildings and Spaces without significant construction works, by an intelligent ICT-based service monitoring and managing the energy consumption. In particular, interoperability between heterogeneous devices and networks, both existing and to be deployed is a key feature to create efficient services and holistic energy control policies. In this paper we describe an innovative software infrastructure to provide a web-service based, hardware independent access to the heterogeneous networks of wireless sensor nodes, such as smart plugs for measuring energy motes for temperature, relative humidity and light monitoring. The proposed infrastructure allows easy extension to other networks, thus representing a contribute to the opening of a market for ICT-based customized solutions integrating numerous products from different vendors and offering services from design of integrated systems to the operation and maintenance phases.

Low-power Embedded System for Real-Time Correction of Fish-Eye Automotive Cameras [p. 340]
M Turturici, S Saponara, L Fanucci and E Franchi

The design and the implementation of a flexible and cost-effective embedded system for real-time correction of fish-eye automotive cameras is presented. Nowadays many car manufacturers already introduced on-board video systems, equipped with fish-eye lens, to provide the driver a better view of the so-called blind zones. A fish-eye lens achieves a larger field of view (FOV) but, on the other hand, causes distortion, both radial and tangential, of the images projected on the image sensor. Since radial distortion is noticeable and dangerous, a real-time system for its correction is presented, whose low-power, low-cost and flexibility features are suitable for automotive applications.

Keywords - Fish-eye camera, video automotive assistance systems, real-time image processing, distortion correction, radial distortion, fish-eye lens, blind zones.

Mechatronic System for Energy Efficiency in Bus Transport [p. 342]
M Donno, A Ferrari, A Scarpelli, P Perlo and A Bocca

Green transport for improving air quality is essential and urgent goal for reaching a healthy environment. In towns with large fleets of public vehicles, technology transfer from standard into new and better solutions requires, in general, time and great investments. This paper presents a quick retrofit for conventional buses in urban transport in order to reduce fuel consumption by using photovoltaic (PV) panels that only recharge the original bus batteries. Experimental tests show that this solution is really effective. Indeed, it could save, per year, several hundred liters of diesel fuel for each bus after considering a solar energy production of about 1.4 MWh.
4.4: System-Level Power and Reliability Estimation and Optimisation


This paper presents an overview of the novel technologies that we are experiencing today in the automation industries. We present the opportunities and challenges of having tightly coupled collaborative networks of embedded systems for controlling complex physical processes. Our objective is to motivate the targeted design automation community to tackle some of the grand challenges in the area of such a distributed, intelligent, and collaborative embedded computing platform.

4.4.1 Intelligent and Collaborative Embedded Computing in Automation Engineering [p. 344]

M A Al Faruque and A Canedo

This paper presents an overview of the novel technologies that we are experiencing today in the automation industries. We present the opportunities and challenges of having tightly coupled collaborative networks of embedded systems for controlling complex physical processes. Our objective is to motivate the targeted design automation community to tackle some of the grand challenges in the area of such a distributed, intelligent, and collaborative embedded computing platform.

4.4.2 Variation-Aware Leakage Power Model Extraction for System-Level Hierarchical Power Analysis [p. 346]

Y Xu, B Li, R Hasholzner, B Rohfleisch, C Haubelt and J Teich

System-level power analysis is commonly used in modern SoC design processes to evaluate power consumption at early design phases. With the increasing variations in manufacturing, the statistical characteristics of parameters are also incorporated in the state-of-the-art methods. However, the spatial correlation between modules still remains as a challenge for system-level statistical power analysis where power models generated from individual modules are used for analysis efficiency or IP protection. In this paper, we propose a novel method to extract variation-aware and correlation-inclusive leakage power models for fast and accurate system-level analysis. For each individual module we generate a power model with different correlation information specified by the module vendor or customer. The local random variables in the power models are replaced by the corresponding ones at system level to reconstruct the correlation between modules so that the accuracy of system-level analysis is guaranteed. Experimental results show that our method are very accurate while being 1000X faster than Monte Carlo simulation and 70X-100X faster than the flattened full chip statistical leakage analysis.

4.4.3 Runtime Power Estimator Calibration for High-Performance Microprocessors [p. 352]

H Wang, S X-D Tan, X-X Liu and A Gupta

Accurate runtime power estimation is important for on-line thermal/power regulation on today’s high performance processors. In this paper, we introduce a power calibration approach with the assistance of on-chip physical thermal sensors. It is based on a new error compensation method which corrects the errors of power estimations using the feedback from physical thermal sensors. To deal with the problem of limited number of physical thermal sensors, we propose a statistical power correlation extraction method to estimate powers for places without thermal sensors. Experimental results on standard SPEC benchmarks show the new method successfully calibrates the power estimator with very low overhead introduced.

4.4.4 Estimation Based Power and Supply Voltage Management for Future RF-Powered Multi-Core Smart Cards [p. 358]

N Druml, C Steger, R Weiss, A Genser and J Haid

RF-powered smart cards are constrained in their operation by their power consumption. Smart card application designers must pay attention to power consumption peaks, high average power consumption and supply voltage drops. If these hazards are not handled properly, the smart card’s operational stability is compromised. Here we present a novel multi-core smart card design, which improves the operational stability of nowadays used smart cards. Estimation based techniques are applied to provide cycle accurate power and supply voltage information of the smart card in real time. A supply voltage management unit monitors the provided power and supply voltage information, flattens the smart card’s power consumption and prevents supply voltage drops by means of a dynamic voltage and frequency scaling (DVFS) policy. The presented multi-core smart card design is evaluated on a hardware emulation platform to prove its proper functionality. Experimental tests show that harmful power variations can be reduced by up to 75% and predefined supply voltage levels are maintained properly. The presented analysis and management functionalities are integrated at a minimal area overhead of 10.1%.

4.4.5 Application-Specific Memory Partitioning for Joint Energy and Lifetime Optimization [p. 364]

H Mahmood, M Poncino, M Loghi and E Macii

Power management of caches based on turning idle cache lines into a low-energy state is also beneficial for the aging effects caused by Negative Bias Temperature Instability (NBTI), provided that idleness is correctly exploited; unlike energy, aging, being a measure of delay, is in fact a worst-case metric. In this work we propose an application-specific partitioned cache architecture in which a cache is organized as a set of independently addressable sub-blocks, by properly using the idleness of the various banks to drive how the partition is determined.
possible to extend the effective lifetime of the cache while saving extra energy. Two are the distinctive features of our approach: First, we
allow the cache sub-blocks age at different rates, achieving a sort of graceful degradation of performance while extending lifetime beyond
the limits of previously published works. Proper architectural arrangements are also introduced in order to cope with the issue of using a
progressively smaller cache. Second, the sub-blocks have non-uniform sizes, so to maximally exploit idleness for joint energy and aging
optimization. Simulation results show that it is possible to extend the effective lifetime of the cache by more than 2x with respect to previous
methods, while concurrently improving energy consumption by about 50%.

4.5: EMBEDDED TUTORIAL - State-of-the-Art Tools and Techniques for Quantitative Modeling and Analysis of Embedded Systems
Moderators: A Legay, INRIA/Rennes, FR

M Bozga, A David, A Hartmanns, H Hermanns, K G Larsen, A Legay and J Tretmans

This paper surveys well-established/recent tools and techniques developed for the design of rigorous embedded systems. We will first survey
UPPAAL and MODEST, two tools capable of dealing with both timed and stochastic aspects. Then, we will overview the BIP framework for
modular design and code generation. Finally, model-based testing will be discussed.

4.6: Compilers and Source-Level Simulation
Moderators: R Rabbah, IBM Research, US; B Franke, Edinburgh U, UK

[89] Hybrid Source-Level Simulation of Data Caches Using Abstract Cache Models [p. 376]
S Stattelmann, G Gebhard, C Cullmann, O Bringmann and W Rosenstiel

This paper presents a hybrid cache analysis for the simulation-based evaluation of data caches in embedded systems. The proposed
technique uses static analyses at the machine code level to obtain information about the control flow of a program and the memory accesses
contained in it. Using the result of these analyses, a high-speed source-level simulation model is generated from the source code of the
application, enabling a fast and accurate evaluation of its data cache behavior. As memory accesses are obtained from the binary-level
control flow, which is simulated in parallel to the original functionality of the software, even complex compiler optimizations can be modeled
accurately. Experimental results show that the presented source-level approach estimates the cache behavior of a program within the same
level of accuracy as established techniques working at the machine code level.
Index Terms - System analysis and design; Timing; Modeling; Software performance; Cache memories;

[90] Accurate Source-Level Simulation of Embedded Software with Respect to Compiler Optimizations [p. 382]
Z Wang and J Henkel

Source code instrumentation is a widely used method to generate fast software simulation models by annotating timing information into
application source code. Source-level simulation models can be easily integrated into SystemC based simulation environment for fast
simulation of complex multiprocessor systems. The accurate backannotation of the timing information relies on the mapping between source
code and binary code. The compiler optimizations might make it hard to get accurate mapping information. This paper addresses the
mapping problems caused by complex compiler optimizations, which are the main source of simulation errors. To obtain accurate mapping
information, we propose a method called fine-grained flow mapping that establishes a mapping between sequences of control flow of source
code and binary code. In case that the code structure of a program is heavily altered by compiler optimizations, we propose to replace the
altered part of the source code with functionally-equivalent IR-level code which has an optimized structure, leading to Partly Optimized
Source Code (POSC). Then the flow mapping can be established between the POSC and the binary code and the timing information is back-
annotated to the POSC. Our experiments demonstrate the accuracy and speed of simulation models generated by our approach.

D She, Y He, B Mesman and H Corporaal

In modern processor architectures, the register file (RF) consumes considerable amount of the processor power. It is well known that by
allowing software to have explicit fine-grained control over the datapath, the transport-triggered architectures (TTAs) can substantially reduce
the RF traffic, thereby minimizing the RF energy. However, it is important to make sure that the gain in RF is not cancelled out by the
overhead due to the fine-grained datapath control, in particular, the deterioration of code density in conventional TTAs. In this paper, we
analyze the potential of minimizing RF energy in MOVE-Pro, a TTA-based processor framework. We present a flexible compiler backend,
which performs energy-aware instruction scheduling to push the limit of RF energy reduction. The experimental results show that with the
proposed energy-aware compiler backend, MOVE-Pro is able to significantly reduce RF energy compared to its RISC/VLIW counterparts, by up to 80%. Meanwhile the code density of MOVE-Pro remains at the same level as its RISC/VLIW counterparts, allowing the energy saving in RF to be successfully transferred to total energy saving.

Index Terms - TTA, MOVE-Pro, Low Power, Code Generation, Register File

D Cordes and P Manwedel

A large amount of research work has been done in the area of automatic parallelization for decades, resulting in a huge amount of tools, which should relieve the designer from the burden of manually parallelizing an application. Unfortunately, most of these tools are only optimizing the execution time by splitting up applications into concurrently executed tasks. In the domain of embedded devices, however, it is not sufficient to look only at this criterion. Since most of these devices are constraint-driven regarding execution time, energy consumption, heat dissipation and other objectives, a good trade-off has to be found to efficiently map applications to multiprocessor system on chip (MPSoC) devices. Therefore, we developed a fully automated multi-objective aware parallelization framework, which optimizes different objectives at the same time. The tool returns a Pareto-optimal front of solutions of the parallelized application to the designer, so that the solution with the best trade-off can be chosen.

Index Terms - Automatic Parallelization, Embedded Software, Multi-Objective, Genetic Algorithms, Task-Level Parallelism, Energy awareness

4.7: Advances in Test Generation

Moderators: G Mrugalski, Mentor Graphics, PL; S Hellebrand, Paderborn U, DE

K-H Chang, H-Z Chou and I L Markov

At-speed testing is increasingly important at recent technology nodes due to growing uncertainty in chip manufacturing. However, aspeed fault coverage and test-efficacy suffer when tests are not robust. Since Automatic Test Pattern Generation (ATPG) is typically performed at late design stages, fixing robustness problems found during ATPG can be costly. To address this challenge, we propose a methodology that identifies robustness problems at the Register Transfer Level (RTL) and fixes them. Empirically, this improves final at-speed fault coverage and test-efficacy.

N Karimi, K Chakrabarty, P Gupta and S Patil

Clock-domain crossing (CDC) faults are a serious concern for high-speed, multi-core integrated circuits. Even when robust design methods based on synchronizers and design verification techniques are used, process variations can introduce subtle timing problems that affect data transfer across clock-domain boundaries for fabricated chips. We present a test generation technique that leverages commercial ATPG tools, but introduces additional constraints, to detect CDC faults. We also present HSpice simulation data using a 45 nm technology to quantify the occurrence of CDC faults at clock-domain boundaries. Results are presented for synthesized IWLS05 benchmarks that include multiple clock domains. The results highlight the ineffectiveness of commercial transition-delay fault ATPG and the “coverage gap” resulting from the use of ATPG methods employed in industry today. While the proposed method can detect nearly all CDC faults, TDF ATPG is found to be severely deficient for screening CDC faults.

D Sabena, M Sonza Reorda and L Sterpone

Feature size reduction drastically influences permanent faults occurrence in nanometer technology devices. Among the various test techniques, Software-Based Self-Test (SBST) approaches have been demonstrated to be an effective solution for detecting logic defects, although achieving complete fault coverage is a challenging issue due to the functional-based nature of this methodology. When VLIW processors are considered, standard processor-oriented SBST approaches result deficient since not able to cope with most of the failures affecting VLIW multiple parallel domains. In this paper we present a novel SBST algorithm specifically oriented to test the register files of VLIW processors. In particular, our algorithm addresses the cross-bar switch architecture of the VLIW register file by completely covering the intrinsic faults generated between the multiple computational domains. Fault simulation campaigns comparing previously developed methods with our solution demonstrate its effectiveness. The results show that the developed algorithm achieves a 97.12% fault coverage which is about twice better than previously developed SBST algorithms. Further advantages of our solution are the limited overhead in terms of execution cycles and memory occupation.

Keywords - Testing, software-based self test, Very Long Instruction Word Processors, Fault Simulation.
Adequate coverage of small-delay defects in circuits affected by statistical process variations requires identification and sensitization of multiple paths through potential defect sites. Existing K longest path generation (KLPG) algorithms use a data structure called path store to prune the search space by restricting the number of sub-paths considered at the same time. While this restriction speeds up the KLPG process, the algorithms lose their optimality and do not guarantee that the K longest sensitizable paths are indeed found. We investigate, for the first time, the effects of missing some of the longest paths on the defect coverage. We systematically quantify how setting different limits on the path-store size affects the numbers and relative lengths of identified paths, as well as the run-times of the algorithm. We also introduce a new optimal KLPG algorithm that works iteratively and pinpointedly addresses defect locations for which the path-store size limit has been exceeded in previous iterations. We compare this algorithm with a naive KLPG approach that achieves optimality by setting the path-store size limit to a very large value. Extensive experiments are reported for 45nm-technology data.

Index Terms - Parameter variations, small-delay testing, K longest path generation
physically active subjects. To validate our experiments, we develop a closed-loop, ambulatory ECG monitoring platform, which consumes 2.17 mW of power and delivers a data rate of 33 kbps over a dedicated UWB link.

Nowadays, sensor networks are being used to monitor increasingly complex physical systems, necessitating advanced signal analysis capabilities as well as the ability to handle large amounts of network data. For the first time, we present a methodology to enable advanced decision support on a low-power sensor node through the direct use of compressively-sensed signals in a supervised-learning framework; such signals provide a highly efficient means of representing data in the network, and their direct use overcomes the need for energy-intensive signal reconstruction. Sensor networks for advanced patient monitoring are representative of the complexities involved. We demonstrate our technique on a patient-specific seizure detection algorithm based on electroencephalograph (EEG) sensing. Using data from 21 patients in the CHB-MIT database, our approach demonstrates an overall detection sensitivity, latency, and false alarm rate of 94.70%, 5.83 seconds, and 0.199 per hour, respectively, while achieving data compression by a factor of 10x. This compares well with the state-of-the-art baseline detector with corresponding results being 96.02%, 4.59 seconds, and 0.145 per hour, respectively.

This paper presents a fully integrated application specific integrated circuit (ASIC) sensor for the recording of multiple bio-electric signals. It consists of an analog front-end circuit with tunable bandwidth and programmable gain, a 6-input 8-bit successive approximation register analog to digital converter (SAR ADC), and a reconfigurable digital core. The ASIC is fabricated in a 0.18-μm 1P6M CMOS technology, occupies an area of 1.5 x 3.0 mm², and totally consumes a current of 16.7 μA from a 1.2 V supply. Incorporated with the ASIC, an Intelligent Electrode can be dynamically configured for on-site measurement of different bio-signals. A 2-wire data transmission protocol is also integrated on chip. It enables the serial connection over a group of Intelligent Electrodes, thus minimizes the number of connecting cables. A wearable healthcare system is built upon a printed Active Cable and a scalable number of Intelligent Electrodes. The system allows synchronous processing of maximum 14-channel bio-signals. The ASIC performance has been successfully verified in in-vivo bio-electric recording experiments.

5.4: Microarchitecture

Branch prediction is critical in exploring instruction level parallelism for modern processors. Previous aggressive branch predictors generally require significant amount of hardware storage and complexity to pursue high prediction accuracy. This paper proposes the Compiler-guided History Stack (CHS), an energy-efficient compiler-microarchitecture cooperative technique for branch prediction. The key idea is to track very-long-distance branch correlation using a low-cost compiler-guided history stack. It relies on the compiler to identify branch correlation based on two program substructures: loop and procedure, and feed the information to the predictor by inserting guiding instructions. At runtime, the processor dynamically saves and restores the global history using a low-cost history stack structure according to the compiler-guided information. The modification on the global history enables the predictor to track very-long-distance branch correlation and thus improves the prediction accuracy. We show that CHS can be combined with most of existing branch predictors and it is especially effective with small and simple predictors. Our evaluations show that the CHS technique can reduce the average branch mispredictions by 28.7% over gshare predictor, resulting in average performance improvement of 10.4%. Furthermore, it can also improve those aggressive perceptron, OGEHL and TAGE predictors.

This work introduces a new branch predictor design that increases the perceived predictor capacity without increasing its delay by using a large virtual second-level table allocated in the second-level caches. Virtualization is applied to a state-of-the-art multi-table branch predictor. We evaluate the design using instruction count as proxy for timing on a set of commercial workloads. For a predictor whose size is determined by access delay constraints, accuracy can be improved by 8.7%. Alternatively, the design can be used to achieve the same accuracy as a non-virtualized design while using 25% less dedicated storage.
5.5: Shared Memory Management in Multicore

Moderators: C Silvano, Polimi, IT; M Berekovic, TU Braunschweig, DE

5.5: Shared Memory Management in Multicore

Energy efficiency is becoming a major constraint in processor designs. Every component of the processor should be reconsidered to reduce wasted energy and area. Prefetching is an important technique for tolerating memory latency. Prefetcher designs have important impact on the energy efficiency of the memory hierarchy. Stride prefetchers require little storage, but cannot handle irregular access patterns. Delta correlation (DC) prefetchers can handle complicated access patterns, but waste storage because of storing multiple miss addresses for a stride pattern. Moreover, DC prefetchers waste the bandwidth and energy of the memory hierarchy because they cannot identify whether an address has been prefetched and generate a large number of redundant prefaches. In this paper, we propose a storage and energy efficient data prefetcher called stride/DC (S/DC) to combine the advantages of stride and DC prefetchers. S/DC uses a pattern prediction table (PPT) which stores two recent miss addresses in each entry to capture stride patterns. PPT avoids recording multiple miss addresses for a stride pattern, and thus improves the storage efficiency. When handling stride patterns, each PPT entry maintains a counter for obtaining the last prefetched address to avoid generating redundant prefetches. When handling other patterns, S/DC compares the new predicted address with earlier generated addresses in the prefetch queue and filters the redundant ones. In addition, to expand the filtering scope, S/DC uses a prefetch filter to store addresses evicted from the prefetch queue. In this way, S/DC reduces the bandwidth requirements and energy consumption of prefetching. Experimental results demonstrate that S/DC achieves comparable performance with only 24% of the storage and reduces 11.46% of the L2 cache energy, as compared to the CZone/DC prefetcher.

An Architecture-Level Approach for Mitigating the Impact of Process Variations on Extensible Processors

In this paper, we present an architecture-level approach to mitigate the impact of process variations on extended instruction set architectures (ISAs). The proposed architecture adds one extra cycle to execute custom instructions (CIs) that violate the maximum allowed propagation delay due to the process variations. Using this method, the parametric yield of manufactured chips will greatly improve. The cost is an increase in the cycle latency of some of the CIs, and hence, a slight performance degradation for the extensible processor architectures. To minimize the performance penalty of the proposed approach, we introduce a new merit function for selecting the CIs during the selection phase of the ISA extension design flow. To evaluate the efficacy of the new selection method, we compare the extended ISAs obtained by this method with those selected based on the worst-case delay. Simulation results reveal that a speedup improvement of about 18% may be obtained by the proposed selection method. Also, by using the proposed merit function, the proposed architecture can improve the speedup about 20.7%.

PCASA: Probabilistic Control-Adjusted Selective Allocation for Shared Caches

Chip Multi-Processors (CMPs) are designed with an increasing number of cores to enable multiple and potentially heterogeneous applications to run simultaneously on the same system. However, this results in increasing pressure on shared resources, such as shared caches. With multiple processor cores sharing the same caches, high-priority applications may end up contending with low-priority applications for cache space and suffer significant performance slow-down, hence affecting the Quality of Service (QoS). In datacenters, Service Level Agreements (SLAs) impose a reserved amount of computing resources and specific cache space per cloud customer. Thus, to meet SLAs, a deterministic capacity management solution is required to control the occupancy of all applications. In this paper, we propose a novel QoS architecture, based on Probabilistic Selective Allocation (PSA), for priority-aware caches. Further, we show that applying a control-theoretic approach (Proportional Integral controller) to dynamically adjust PSA provides accurate and fine-grained capacity management.

Dynamic Directories: A Mechanism for Reducing On-Chip Interconnect Power in Multicores

On-chip interconnection networks consume a significant fraction of the chip's power, and the rapidly increasing core counts in future technologies is going to further aggravate their impact on the chip's overall power consumption. A large fraction of the traffic originates from data messages exchanged between sharing cores, but from the communication between the cores and intermediate hardware structures (i.e., directories) for the purpose of maintaining coherence in the presence of conflicting updates. In this paper, we propose Dynamic Directories, a method allowing the directories to be placed arbitrarily in the chip by piggy-backing the virtual to physical address translation. This eliminates a large fraction of the on-chip interconnect traversals, hence reducing the power consumption. Through trace-driven and cycle-accurate simulation in a range of scientific and Map-Reduce applications, we show that our technique reduces the power and energy
expended by the on-chip interconnect by up to 37% (16.4% on average) with negligible hardware overhead and a small improvement in performance (1.3% on average).

Keywords-On-chip networks; Non-uniform caches; Multicore architecture

Non-Uniform Cache Access (NUCA) architectures provide a potential solution to reduce the average latency for the last-level-cache (LLC), where the cache is organized into per-core local and remote partitions. Recent research has demonstrated the benefits of cooperative cache sharing among local and remote partitions. However, ignoring cache access patterns of concurrently executing applications sharing the local and remote partitions can cause inter-partition contention that reduces the overall instruction throughput. We propose a dynamic cache management scheme for LLC in NUCA-based architectures, which reduces inter-partition contention. Our proposed scheme provides efficient cache sharing by adapting migration, insertion, and promotion policies in response to the dynamic requirements of the individual applications with different cache access behaviors. Our adaptive cache management scheme allows individual cores to steal cache capacity from remote partitions to achieve better resource utilization. On average, our proposed scheme increases the performance (instructions per cycle) by 28% (minimum 8.4%, maximum 75%) compared to a private LLC organization.

Barrier synchronization is a key programming primitive for shared memory embedded MPSoCs. As the core count increases, software implementations cannot provide the needed performance and scalability, thus making hardware acceleration critical. In this paper we describe an interconnect extension implemented with standard cells and with a mainstream industrial toolflow. We show that the area overhead is marginal with respect to the performance improvements of the resulting hardware-accelerated barriers. We integrate our HW barrier into the OpenMP programming model and discuss synchronization efficiency compared with traditional software implementations.

In real-time systems, there are two distinct trends for scheduling task sets on unicore systems: non-preemptive and preemptive scheduling. Non-preemptive scheduling is obviously not subject to any preemption delay but its schedulability may be quite poor, whereas fully preemptive scheduling is subject to preemption delay, but benefits from a higher flexibility in the scheduling decisions. The time-delay involved by task preemptions is a major source of pessimism in the analysis of the task Worst-Case Execution Time (WCET) in real-time systems. Preemptive scheduling policies including non-preemptive regions are a hybrid solution between non-preemptive and fully preemptive scheduling paradigms, which enables to conjugate both world's benefits. In this paper, we exploit the connection between the progression of a task in its operations, and the knowledge of the preemption delays as a function of its progression. The pessimism in the preemption delay estimation is then reduced in comparison to state of the art methods, due to the increase in information available in the analysis.

This paper presents a new semi-partitioned approach to schedule sporadic tasks on multi-core platform based on the Rate Monotonic Scheduling (RMS) policy. Our approach exploits the well known fact that harmonic tasks have better schedulability than non-harmonic ones on a single processor. The challenge for our approach, however, is how to take advantage of this fact to assign and split appropriate tasks on different processors in the semi-partitioned approach. We formally prove that our scheduling approach can successfully schedule any task sets with system utilizations bounded by the Liu&Layland's bound. Our extensive experiment results demonstrate that the proposed algorithm can significantly improve the scheduling performance compared with the previous work.

Time-Triggered Network-on-Chip (TTNoC) is a networking concept aiming at providing both predictable and high-throughput communication for modern multiprocessor systems. The message scheduling is one of the major design challenges in TTNoC-based systems. The designers
not only need to allocate time slots but also have to assign communication routes for all messages. This paper tackles the TTNoC scheduling problem and presents an approach based on Satisfiability Modulo Theories (SMT) solving. We first formulate the complete problem as an SMT instance, which can always compute a feasible solution if exists. Thereafter, we propose an incremental approach that integrates SMT solving into classical heuristic algorithms. The experimental results show that the heuristic scales significantly better with only minor loss of performance.

S Quinton, M Hanke and R Ernst

This paper presents a new compositional approach providing safe quantitative information about real-time systems. Our method is based on a new model to describe sporadic overload at the input of a system. We show how to derive from such a model safe quantitative information about the response time of each task. Experiments demonstrate the efficiency of this approach on a real-life example. In addition we improve the state of the art in compositional performance analysis by introducing execution time models which take into account several consecutive executions and by using tighter bounds for computing output event models.

5.7: Testing of Non-Volatile Memories

Moderators: R Aitken, ARM, US; B Tasic, NXP Semiconductors, NL

[114] Error Patterns in MLC NAND Flash Memory: Measurement, Characterization, and Analysis [p. 521]
Y Cai, E F Haratsch, O Mutlu and K Mai

As NAND flash memory manufacturers scale down to smaller process technology nodes and store more bits per cell, reliability and endurance of flash memory reduce. Wear-leveling and error correction coding can improve both reliability and endurance, but finding effective algorithms requires a strong understanding of flash memory error patterns. To enable such understanding, we have designed and implemented a framework for fast and accurate characterization of flash memory throughout its lifetime. This paper examines the complex flash errors that occur at 30-40nm flash technologies. We demonstrate distinct error patterns, such as cycle-dependency, location-dependency and value-dependency, for various types of flash operations. We analyze the discovered error patterns and explain why they exist from a circuit and device standpoint. Our hope is that the understanding developed from this characterization serves as a building block for new error tolerance algorithms for flash memory.

Keywords - NAND flash; error patterns; endurance; reliability; error correction

J Zha, X Cui and C L Lee

Advance of the fabrication technology has enhanced the size and density for the NAND Flash memory but also brought new types of defects which need to be tested for the quality consideration. This work analyzes three types of physical defects for the deep nano-meter NAND Flash memory based on the circuit level simulation and proposes new categories of interference faults (IFs). Testing algorithm is also proposed to test the faults under the worst case condition. The algorithm, in addition to test IFs, can also detect the conventional address faults, disturbance faults and other RAM-like faults for the NAND Flash.

Keywords - NAND Flash; Fault Model; Interference Fault


Magnetic Random Access Memory (MRAM) is an emerging technology with the potential to become the universal on-chip memory. Among the existing MRAM technologies, the Thermally Assisted Switching (TAS) MRAM technology offers several advantages compared to the others technologies: selectivity, single magnetic field and integration density. As any other types of memory, TAS-MRAMs are prone to defects, so TAS-MRAM testing needs definitely to be investigated since only few papers can be found in the literature. In this paper we analyze the impact resistive-open defects on the heat current of a TAS-MRAM architecture. Electrical simulations were performed on a hypothetical 4x4 TAS-MRAM architecture enabling any read/write operations. Results show that W0 and/or W1 operations may be affected by the resistive-open defects. This study provides insights into the various types of TAS-MRAM defects and their behavior. As future work, we plan to utilize these analyses results to guide the test phase by providing effective test algorithm targeting fault related to actual defects that may affect TAS-MRAM architecture.

Keywords - non-volatile memories, spintronics, TAS-MRAM, heat current, resistive-open defects, fault modeling, test.
Aggregate scheduling in routers merges several flows into one aggregate flow. We propose an approach for computing the end-to-end delay bound of individual flows in a FIFO multiplexer under aggregate scheduling. A synthetic case study exhibits that the end-to-end delay bound is up to 33.6% tighter than the case without considering the traffic peak behavior.

On-chip interconnection networks simplify the integration of complex system-on-chips. The switches are the basic building blocks of such networks and their design critically affects the performance of the whole system. The transfer of data between the inputs and the outputs of the switch is performed by the crossbar, whose active connections are decided by the arbiter. In this paper, we design scalable dynamic-priority arbiters that are merged with the crossbar's multiplexers. The proposed RTL macros can adjust to various priority selection policies, while still following the same unified architecture. With this approach, sophisticated arbitration policies that yield significant network-throughput benefits can be implemented with negligible delay cost relative to the standard round-robin policy.

The degradation of CMOS devices over the lifetime can cause the severe threat to the system performance and reliability at deep submicron semiconductor technologies. The negative bias temperature instability (NBTI) is among the most important sources of the aging mechanisms. Applying the traditional guardbanding technique to address the decreased speed of devices is too costly. Due to presence of the narrow-width values, integer register files in high-performance microprocessors suffer a very high NBTI stress. In this paper, we propose an aging-aware register file (AARF) design to combat the NBTI-induced aging in integer register files. The proposed AARF design can mitigate the negative aging effects by balancing the duty cycle ratio of the internal bits in register files. By gating the leading bits of the narrow-width values during the register accesses, our AARF can also achieve a significantly power reduction, which will further reduce the temperature and NBTI degradation of integer register files. Our experimental results show that AARF can effectively reduce the NBTI stress with a 36.9% power saving for integer register files.

Power has become the overriding concern for most modern electronic applications today. To reduce clock power, sequential clock gating is increasingly getting used over and above combinational clock gating. Given the complexity of manually identifying sequential clock gating changes, automatic tools are becoming popular. However, since these tools always work within the scope of the design and the constraints provided, they do not provide any insight into additional power savings that might still be possible. In this paper we present an interactive sequential analysis flow, PowerAdviser, which besides performing automatic sequential changes also provides information for additional power savings that the user can realize through manual changes. Using this new flow we have achieved dynamic power reduction up to 45% more than a purely automated flow.

Towards parallel execution of industrial cyber-physical systems (CPS), the ability of a system to react quicker to its inputs by just a few milliseconds can be translated to billions of dollars in additional profit over just a few years of uninterrupted operation. Therefore, it is important to reduce the cycle time of industrial CPS applications not only for the economical benefits but also for waste minimization, energy reduction, and safer working environments. In this paper, we present a novel method to reduce the execution time of CPS applications through a holistic software/hardware method that enables automatic parallelization of standardized industrial automation languages and their execution in multi-core processors. Through a realistic CPS, we demonstrate that parallel execution reduces the cycle time of the application and increases the life-cycle through better utilization of the mechanical, electrical, and computing resources.
In this paper, we propose an implication graph based sequential logic simulator for debugging scan pattern failures encountered during First Silicon. A novel Debug Implication Graph (DIG) is constructed during logic simulation of the failing scan pattern. An efficient node traversal mechanism across time frames, in the DIG, is used to perform the root-cause analysis for the failing scan-cells. We have developed an Interactive Pattern Debug environment (IDE), viz. scan pattern debugger, around the logic simulator to systematically analyze and root-cause the failures. We integrated the proposed technique into the scan ATPG flow for industrial microprocessor designs. We were able to resolve the First Silicon logical pattern failures within hours, which would have otherwise taken a few days of manual effort.

This paper presents FAST-GP, a framework for functional verification of RTL designs, which is based on fault injection and parallel simulation on GP-GPUs. Given a fault model, the framework translates the RTL code into an injected C code targeting NVIDIA GPUs, thus allowing a very fast parallel automatic test pattern generation and fault simulation. The paper compares different configurations of the framework to better exploit the architectural characteristics of such GPGPUs (such as thread synchronization, branch divergence, etc.) by considering the architectural characteristics of the RTL design under verification (i.e., complexity, size, number of injected faults, etc.). Experimental results have been conducted by applying the framework to different designs, in order to prove the methodology effectiveness.

Complex Application Specific Instruction-set Processors (ASIPs) expose to the designer a large number of degrees of freedom, posing the need for highly accurate and rapid simulation environments. FPGA-based emulators represent an alternative to software cycle-accurate simulators, preserving maximum accuracy and reasonable simulation times. The work presented in this paper aims at exploiting FPGA emulation within technology aware design space exploration of ASIPs. The potential speedup provided by reconfigurable logic is reduced by the overhead of RTL synthesis/implementation. This overhead can be mitigated by reducing the number of FPGA implementation processes, through the adoption of binary-level translation. Hereby we present a prototyping method that, given a set of candidate ASIP configurations, defines an overdimensioned ASIP architecture, capable of emulating all the design space points under evaluation. This approach is then evaluated with a design space exploration case study. Along with execution time, by coupling FPGA emulation with activity-based physical modeling, we can extract area/power/energy figures.

This paper identifies that many WSN applications employ algorithms which can be solved by using parallel prefix-sums. Therefore, an alternative architecture is proposed to calculate them energy-efficiently. It consists of several parallel processing elements (PEs) structured as a folded tree. Profiling SystemC models of the design with ActivaSC helps to improve data-locality. Measurements of the fabricated chip confirm an improvement of 10-20x in terms of energy as compared with traditional MCUs found in sensor nodes.

This paper introduces a power efficient approach for improving reliability of heterogeneous register files in embedded processors. The approach is based on the fact that control applications have high demands in reliability, while many special-purpose register are unused in a considerable portion of execution. The paper proposes a static application binary analysis which is applied at function-level granularity and offers a systematic way to manage the RF's protection by mirroring the content of used registers into unused ones. The simulation results on an enhanced Blackfin processor demonstrate that Register File Vulnerability Factor (RFVF) is reduced from 35% to 6.9% in cost of 1% performance lost on average for control applications from Mibench suite.

Target sensitive tasks have an execution window for feasibility and must execute at a target point in time for maximum utility. In the gravitational task model, a task can express a target point, and the utility decay as a function of the deviation from this point. A method called...
equilibrium approximates the schedule with maximum utility accrual based on an analogy with physical pendulums. In this paper, we propose a scheduling algorithm for this task model to schedule periodic tasks. The basic idea of our solution is to combine the equilibrium with Earliest Deadline First (EDF) in order to reuse EDF's well studied timeliness analysis. We present simulation results and an example multimedia application to show the benefits of our solution.

[128] Online Scheduling for Multi-Core Shared Reconfigurable Fabric [p. 582]
L Chen, T Marconi and T Mitra

Processor customization in the form of application-specific instructions has become a popular choice to meet the increasing performance demands of embedded applications under short time-to-market constraints. Implementing the custom instructions in reconfigurable logic provides greater flexibility. Recently, a number of architectures have been proposed where multiple cores on chip share a single reconfigurable fabric that implements the custom instructions. Effective exploitation of this reconfigurable fabric requires runtime scheduling of the tasks on the cores and allocation of reconfigurable logic for custom instructions. In this paper, we propose an efficient online scheduling algorithm for multi-core shared reconfigurable fabric and show its effectiveness through experimental evaluation.

A Mohammadi, M Ebrahimi, A Ejilali and S G Miremadi

In this paper, we have proposed a fast and easy-to-develop FPGA-based fault injection technique. This technique uses the Altera FPGAs debugging facilities in order to inject SEU fault model in both flip-flops and memory units. Since this method uses the FPGAs built-in facilities, it imposes a negligible performance and area overhead on the system. The experimental results on Leon2 processor shows that the proposed technique is on average four orders of magnitude faster than a simulation-based fault injection.

6.1: PANEL - Role of EDA in the Development of Electric Vehicles (Special Day E-Mobility)

Moderator: O Bringmann, FZI Research Center for Information Technology, Karlsruhe, DE

Electric vehicles will only get widely accepted if driving range, comfort and safety do not differ significantly from today's cars with internal combustion engine. Microelectronics will play a remarkable role in implementing e.g. optimized energy management systems using situation-aware recuperation strategies, smart (re-)charging capabilities and advanced driving and operation strategies in upcoming EVs. However, the development process has to be closely interlinked across different domains which results into many new challenges for the EDA community. Therefore, this panel will discuss visions and recent advances within an interdisciplinary field of competence by bringing together leading tool vendors from different domains.

6.1.2: Keynote

[131] Research and Innovation on Advanced Computing - an EU Perspective [p. 591]
T Van der Pyl, Director Components and Systems, European Commission

Under 'Components and Systems' in FP7-ICT, over the period 2007-2012, the EU has so far invested about 100M on Computing Systems research. Building on the industrial constituencies and activities of the Joint Technology Initiative ARTEMIS and complementing research on embedded systems and control, research and innovation on Computing Systems covers a broad spectrum of issues from multi-core scalability and mastering parallelism to hardware/software co-design and low energy/low cost chips. With the convergence of computing technologies, work covers the broad spectrum of computing systems from customised computing via data servers to high performance systems. Work builds on and expands from European industrial strengths in embedded and mobile computing with low cost and energy efficiency being key drivers. After a short overview of the research supported, some major trends in computing systems and their role in our society will be discussed. First ideas of new funding opportunities under Advanced Computing Systems in ICT Work Programme 2013 will be outlined. An outlook towards the next Framework Programme for Research and Innovation "Horizon 2020" and an overview of recommendations received from consultation activities with the constituencies in the broad context of Computing will conclude the presentation.

6.2: EMBEDDED TUTORIAL - Memristor Technology

Moderator: R Tetzlaff, TU Dresden, DE

R Tetzlaff, A Bruening, L O Chua, R S Williams
6.3: Thermal Aware Low Power Design

Moderators: A Macii, Politecnico di Torino, IT; A Garcia-Ortiz, Bremen U, DE

Thermal Aware Low Power Design

Mapping of memory accesses to DRAM banks also present a memory management policy that targets applications with spatial variations in DRAM accesses and performs temperature-aware systems running parallel applications demonstrate up to 88.5% improvement in energy delay product compared to equivalent 2D systems. We simulations along with energy and thermal evaluations, including a detailed analysis of the DRAM layers. Experimental results on 16-core 3D evaluation framework. This paper first presents such a framework for 3D multicore systems capable of running architecture-level performance interplay between performance, energy, and temperature on 3D systems with on-chip DRAM can only be addressed using a comprehensive slow off-chip buses. Power densities and temperatures on the chip, however, increase following the performance improvement. The complex 3D stacked systems with on-chip DRAM provide high speed and wide bandwidth for accessing main memory, overcoming the limitations of slow off-chip buses. Power densities and temperatures on the chip, however, increase following the performance improvement. The complex interplay between performance, energy, and temperature on 3D systems with on-chip DRAM can only be addressed using a comprehensive evaluation framework. This paper first presents such a framework for 3D multicore systems capable of running architecture-level performance simulations along with energy and thermal evaluations, including a detailed analysis of the DRAM layers. Experimental results on 16-core 3D systems running parallel applications demonstrate up to 88.5% improvement in energy delay product compared to equivalent 2D systems. We also present a memory management policy that targets applications with spatial variations in DRAM accesses and performs temperature-aware mapping of memory accesses to DRAM banks.

---

The memristor is a new nano-electronic device very promising for emerging technologies. Although 40 years ago Leon Chua has postulated this circuit element, only the invention of the crossbar latch by the HP group of Stanley Williams provided the first nanoelectronic realization of such a device in 2008. Thus it has been shown that the ideal circuit elements (R,C,L) were not sufficient to model basic real-world circuits. Memristors being essentially resistors with memory are able to perform logic operations as well as storage of information. Recently, it has been announced that “Williams expects to see memristors used in computer memory chips within the next few years. HP Labs already has a production-ready architecture for such a chip” (http://www.hpl.hp.com/news/2010/aprjun/memristor.html [133]). Memristors are outstanding candidates for future analog, digital, and mixed signal circuits.

---

6.3: Thermal Aware Low Power Design

Moderators: A Macii, Politecnico di Torino, IT; A Garcia-Ortiz, Bremen U, DE

[PDF] TempoMP: Integrated Prediction and Management of Temperature in Heterogeneous MPSoCs [p. 593]
S Shanfi, R Ayoub and T Srinic Rosing

Heterogeneous Multi-Processor Systems on a Chip (MPSoCs) are more complex from a thermal perspective compared to the homogeneous MPSoCs because of their inherent imbalance in power density. In this work we develop TempoMP, a new technique for thermal management of heterogeneous MPSoCs which leverages multi-parametric optimization along with our novel thermal predictor, Tempo. TempoMP is able to deliver locally optimal dynamic thermal management decisions to meet thermal constraints while minimizing power and maximizing performance. It leverages our Tempo predictor which, unlike the previous techniques, can estimate the impact of future power state changes at negligible overhead. Our experiments show that compared to the state of the art, Tempo can reduce the maximum prediction error by up to an order of magnitude. Our experiments with heterogeneous MPSoCs also show that TempoMP meets thermal constraints while reducing the average task lateness by 2.5X and energy-lateness product by 5X compared to the state of the art techniques.

[PDF] Thermal Balancing of Liquid-Cooled 3D-MPSoCs Using Channel Modulation [p. 599]
M M Sabry, A Sridhar and D Atienza

While possessing the potential to replace conventional air-cooled heat sinks, inter-tier microchannel liquid cooling of 3D ICs also creates the problem of increased thermal gradients from the fluid inlet to outlet ports [1, 2]. These cooling-induced thermal gradients can be high enough to create undesirable stress in the ICs, undermining the structural reliability and lifetimes. In this paper, we present a novel design-time solution for the thermal gradient problem in liquid-cooled 3D Multi-Processor System-on-Chip (MPSoC) architectures. The proposed method is based on channel width modulation and provides the designers with an additional dimension in the design-space exploration. We formulate the channel width modulation as an optimal control design problem to minimize the temperature gradients in the 3D IC while meeting the design constraints. The proposed thermal balancing technique uses an analytical model for forced convective heat transfer in microchannels, and has been applied to a two tier 3D-MPSoC. The results show that the proposed approach can reduce thermal gradients by up to 31% when applied to realistic 3D-MPSoC architectures, while maintaining pressure drops in the microchannels well below their safe limits of operation.

D-C Juan, Y-L Chuang, D Marculescu, Y-W Chang

Unaddressed thermal issues can seriously hinder the development of reliable and low power systems. In this paper, we propose a statistical approach for analyzing thermal behavior under leakage power variations stemming from the manufacturing process. Based on the proposed models, we develop floorplanning techniques targeting thermal optimization. The experimental results show that peak temperature is reduced by up to 8.8°C, while thermal-induced leakage power and maximum thermal variance are reduced by 13% and 17%, respectively, with no additional area overhead compared with best performance-driven optimized design.

J Meng and A K Coskun

3D stacked systems with on-chip DRAM provide high speed and wide bandwidth for accessing main memory, overcoming the limitations of slow off-chip buses. Power densities and temperatures on the chip, however, increase following the performance improvement. The complex interplay between performance, energy, and temperature on 3D systems with on-chip DRAM can only be addressed using a comprehensive evaluation framework. This paper first presents such a framework for 3D multicore systems capable of running architecture-level performance simulations along with energy and thermal evaluations, including a detailed analysis of the DRAM layers. Experimental results on 16-core 3D systems running parallel applications demonstrate up to 88.5% improvement in energy delay product compared to equivalent 2D systems. We also present a memory management policy that targets applications with spatial variations in DRAM accesses and performs temperature-aware mapping of memory accesses to DRAM banks.
A Guiding Coverage Metric for Formal Verification

F Haedicke, D Grosse and R Drechsler

Considerable effort is made to verify the correct functional behavior of circuits and systems. To guarantee the overall success, metric-driven verification flows have been developed. In these flows, coverage metrics are omnipresent. Well-established coverage metrics for simulation-based verification approaches exist. This is however not the case for formal verification where property checking is a major technique to prove the correctness of the implementation. In this paper, we present a guiding coverage metric for this formal verification setting. Our metric reports a single number describing how much of the circuit behavior is uniquely determined by the properties. In addition, the coverage metric guides the verification engineer to achieve completeness by providing helpful information about missing scenarios. This information comes from a new behavior classification algorithm that determines uncovered behavior classes for a signal and allows to compute the coverage of a signal. To measure the complete circuit behavior, we devise a coverage metric for a set of signals. The metric is calculated by partitioning the coverage computation into a safe part and an unsafe part, where the latter is weighted accordingly using recursion. This procedure takes into account that in practice, properties refer to internal signals which in turn need to be covered themselves. Overall, our metric allows to track the verification progress in property checking and significantly aid the verification engineers in completing the property set.

Verification of Partial Designs Using Incremental QBF Solving

P Marin, C Miller, M Lewis and B Becker

SAT solving is an indispensable core component of numerous formal verification tools and has found widespread use in industry, in particular when using it in an incremental fashion, e.g., in Bounded Model Checking (BMC). On the other hand, there are applications, in particular in the area of partial design verification, where SAT formulas are not expressive enough and a description via Quantified Boolean Formulas (QBF) is much more adequate. In this paper, we introduce incremental QBF solving and thereby make it usable as a core component of BMC. To do so, we realized an incremental version of the state-of-the-art QBF solver QuBE, allowing for the reuse of learnt information, e.g., in the form of conflict clauses and solution cubes. As an application, we consider BMC for partial designs (i.e., designs containing so-called blackboxes) and thereby disprove realizability, that is, we prove that an unsafe state is reachable no matter how the blackboxes are implemented. In our experimental analysis, we compare different incremental approaches implemented in our BMC tool. BMC with incremental QBF turns out to be feasible for designs with more than 21,000 gates and 2,700 latches. Significant performance gains over non-incremental QBF-based BMC can be obtained on many benchmark circuits, in particular when using the so-called backward-incremental approach combined with incremental preprocessing.

Non-Solution Implications Using Reverse Domination in a Modern SAT-based Debugging Environment

B Le, H Mangassarian, B Keng and A Veneris

With the growing complexity of VLSI designs, functional debugging has become a bottleneck in modern CAD flows. To alleviate this cost, various SAT-based techniques have been developed to automate bug localization in the RTL. In this context, dominance relationships between circuit blocks have been recently shown to reduce the number of SAT solver calls, using the concept of solution implications. This paper first introduces the dual concepts of reverse domination and non-solution implications. A SAT solver is tailored to leverage reverse dominators for the early on-the-fly detection of bug-free components. These are non-solution areas and their early pruning significantly reduces the debugging search-space. This process is expedited by branching on error-select variables first. Extensive experiments on tough real-life industrial debugging cases show an average speedup of 1.7x in SAT solving time over the state-of-the-art, a testimony of the practicality and effectiveness of the proposed approach.

Optimizing Performance Analysis for Synchronous Dataflow Graphs with Shared Resources

D Thiele and R Ernst

Contemporary embedded systems, which process streaming data such as signal, audio, or video data, are an increasingly important part of our lives. Shared resources (e.g., memories) help to reduce the chip area and power consumption of these systems, saving costs in high-volume consumer products. Resource sharing, however, introduces new timing interdependencies between system components, which must be analyzed to verify that the initial timing requirements of the application domain are still met. Graphs with synchronous dataflow (SDF)
semantics are frequently used to model these systems. In this paper, we present a method to integrate resource sharing into SDF graphs. Using these graphs and a throughput constraint, we will derive deadlines for resource accesses and the amount of memory required for an implementation. Then we derive the resource load directly from the SDF description, and perform a formal schedulability analysis to check if the original timing constraints are still met. Finally, we perform an evaluation of our approach using an image processing application and present our results.

Compositional System-Level Design Exploration with Planning of High-Level Synthesis [p. 641]
H-Y Liu, M Petracca and L P Carloni

The growing complexity of System-on-Chip (SoC) design calls for an increased usage of transaction-level modeling (TLM), high-level synthesis tools, and reuse of pre-designed components. In the framework of a compositional methodology for efficient SoC design exploration we present three main contributions: a concise library format for characterization and reuse of components specified in high-level languages like SystemC; an algorithm to prune alternative implementations of a component given the context of a specific SoC design; and an algorithm that explores compositionally the design space of the SoC and produces a detailed plan to run high-level synthesis on its components for the final implementation. The two algorithms are computationally efficient and enable an effective parallelization of the synthesis runs. Through a case study, we show how our methodology returns the essential properties of the design space at the system level by combining the information from the library of components and by identifying automatically those having the most critical impact on the overall design.

Correct-by-Construction Multi-Component SoC Design [p. 647]
R Sinha, P S Roop, Z Salcic and S Basu

Systems-on-chip (SoCs) contain multiple interconnected and interacting components. In this paper, we present a compositional approach for the integration of multiple components with a wide range of protocol mismatches into a single SoC. We show how SoC construction can be done in single-step when all components are integrated at once or it can also be performed incrementally by adding components to an already integrated design. Using a number of AMBA IPs, we show that the proposed framework is able to perform protocol conversion in many cases where existing approaches fail.

Model Checking of Scenario-Aware Dataflow with CADP [p. 653]
B Theelen, J-P Katoen and H Wu

Various dataflow formalisms have been used for capturing the potential parallelism in streaming applications to realise distributed (multi-core) implementations as well as for analysing key properties like absence of dead-lock, throughput and buffer occupancies. The recently introduced formalism of Scenario-Aware Dataflow (SADF) advances these abilities by appropriately capturing the dynamism in modern streaming applications like MPEG-4 video decoding. This paper reports on the application of Interactive Markov Chains (IMC) to capture SADF and to formally verify functional and performance properties. To this end, we propose a compositional IMC semantics for SADF based on which the Construction and Analysis of Distributed Processes (CADP) tool suite enables model checking various properties. Encountered challenges included dealing with probabilistic choice and potentially unbounded buffers, both of which are not natively supported, as well as a fundamental difference in the underlying time models of SADF and IMC. Application of our approach to an MPEG-4 decoder revealed state space reduction factors up to about 21 but also some limitations in terms of scalability and the performance properties that could be analysed.

An Instruction Scratchpad Memory Allocation for the Precision Timed Architecture [p. 659]
A Prakash and H D Patel

This work presents a static instruction allocation scheme for the precision timed architecture's (PRET) scratchpad memory. Since PRET provides timing instructions to control the temporal execution of programs, the objective of the allocation scheme is to ensure that the explicitly specified temporal requirements are met. Furthermore, this allocation incorporates instructions from multiple hardware threads of the PRET architecture. We formulate the allocation as an integer-linear programming problem, and we implement a tool that takes binaries, constructs a control-flow graph, performs the allocation, rewrites the binary with the new allocation, and generates an output binary for the PRET architecture. We carry out experiments on a subset of a modified version of the Malardalen benchmarks to show the benefits of performing the allocation across multiple threads.

Bounding WCET of Applications Using SDRAM with Priority Based Budget Scheduling in MPSoCs [p. 665]
H Shah, A Raabe and A Knoll
SDRAM is a popular off-chip memory that provides large data storage, high data rates, and is in general significantly cheaper than SRAM. There is a growing interest in using SDRAMs in safety critical application domains like aerospace, automotive and industrial automation. Some of these applications have hard real-time requirements where missing a deadline can have devastating consequence. Before integrating any hardware or software in this type of system it needs to be proven that deadlines will always be met. In practice, this is done by analyzing application's timing behavior and calculating its Worst Case Execution Time (WCET). SDRAMs have variable access latencies depending on the refresh operation and the previous accesses. This paper builds on hardware techniques such as bank interleaving and applying Priority Based Budget Scheduling (PBS) to share the SDRAM among multiple masters. Its main contribution is a technique to bound the WCET of an application accessing a shared SDRAM of a multicore architecture using the worst case access pattern. We implemented and tested an overall memory system on an Altera Cyclone III FPGA and applied the proposed WCET estimation technique. The results show that our technique produces safe and low WCET bounds.

M Gerdes, F Kluge, T Ungerer, C Rochange and P Sainrat

In this paper we present synchronisation techniques for hard real-time (HRT) capable execution of parallelised applications on embedded multi-core processors. We show how commonly used software synchronisation techniques can be implemented in a time analysable way based on the proposed hardware primitives. We choose to implement the hardware synchronisation primitives in the memory controller for two reasons. Firstly, we remove pessimism in the WCET analysis of parallelised HRT applications. Secondly, we enable that the implementation of synchronisation techniques is mostly independent of the chosen instruction set architecture (ISA) which allows to use the existing ISAs without enhancements. We analyse the presented synchronisation techniques with the static worst-case execution time (WCET) analysis tool OTAWA. In summary, our specifically engineered synchronisation techniques yield a tremendous gain on the WCET of parallelised HRT applications.


This session brings together specialists from the DfT, DfY and DfR domains that will address key problems together with their solutions for the 14nm node and beyond, dealing with extremely complex chips affected by high defect levels, unpredictable and heterogeneous timing behavior, circuit degradation over time, including extreme situations related with the ultimate CMOS nodes, where all processor nodes, routers and links of single-chip massively parallel tera-device processors could comprise timing faults (such as delay faults or clock skews); a large percentage of these parts are affected by catastrophic failures; all parts experience significant performance degradations over time; and new catastrophic failures occur at low MTBF.
Keywords: DfT, DfY, DfR, ultimate CMOS, single-chip massively parallel teradevice processors.

K Knoedler, J Steinmann, S Laversanne, S Jones, A Huss, E Kural, O Sanchez, O Bringmann, J Zimmermann

This paper briefly describes the latest achievements of a new functional vehicle system to overcome the range anxiety problem of Fully Electric Vehicles (FEV). This is primarily achieved by integrated control and operation strategies to optimize the driving range. The main focus of these control strategies is cooperated electric drivetrain and regenerative braking system. The diverse source of information with on-board and off-board sensors, including navigation system, satellite information, car-to-car, car-to-infrastructure communication and radar and camera systems are primarily utilized to maximize the energy efficiency and correspondingly the range of the FEV.
Keywords: Energy manager, FEV range anxiety, all electric range, network architecture, control, operation strategies, vehicle simulation, regenerative vacuum free braking, environmental sensors, radar, video, satellite navigation, safety, GPS, car-to-car, car-to-infrastructure, Hybrid Electric Vehicles (HEV).
7.3: Multimedia and Consumer Applications

The case for developing and using virtual platforms (VPs) has now been made. If developers of complex HW/SW systems are not using VPs for their current design, complexity of next generation designs demands for their adoption. In addition, the users of these complex systems are asking either for virtual or real platforms in order to develop and validate the software that runs on them, in context with the hardware that is used to deliver some of the functionality. Debugging the erroneous interactions of events and state in a modern platform when things go wrong is hard enough on a VP; on a real platform (such as an emulator or FPGA-based prototype) it can become impossible unless a new level of sophistication is offered. The priority now is to ensure that the capabilities of these platforms meet the requirements of every application domain for electronics and software-based product design. And to ensure that all the use cases are satisfied. A key requirement is to keep pace with Moore’s Law and the ever increasing embedded SW complexity by providing novel simulation technologies in every product release. This paper summarizes a special session focused on the latest applications and latest use cases for VPs. It gives an overview of where this technology is going and the impact on complex system design and verification.

7.3.1 Multimedia and Consumer Applications

Moderators: T Theocharides, Cyprus U, CY; F Kienle, TU Kaiserslautern, DE

7.3.1.1 Virtual Platforms: Breaking New Grounds [p. 685]
R Leupers, G Martin, R Plyaskan, A Herkersdorf, F Schirmeister, T Kogel, M Vaupel

The case for developing and using virtual platforms (VPs) has now been made. If developers of complex HW/SW systems are not using VPs for their current design, complexity of next generation designs demands for their adoption. In addition, the users of these complex systems are asking either for virtual or real platforms in order to develop and validate the software that runs on them, in context with the hardware that is used to deliver some of the functionality. Debugging the erroneous interactions of events and state in a modern platform when things go wrong is hard enough on a VP; on a real platform (such as an emulator or FPGA-based prototype) it can become impossible unless a new level of sophistication is offered. The priority now is to ensure that the capabilities of these platforms meet the requirements of every application domain for electronics and software-based product design. And to ensure that all the use cases are satisfied. A key requirement is to keep pace with Moore’s Law and the ever increasing embedded SW complexity by providing novel simulation technologies in every product release. This paper summarizes a special session focused on the latest applications and latest use cases for VPs. It gives an overview of where this technology is going and the impact on complex system design and verification.

7.3.1.2 An FPGA-based Accelerator for Cortical Object Classification [p. 691]
M S Park, S Kestur, J Sabarad, V Narayanan and M J Irwin

Recently significant advances have been achieved in understanding the visual information processing in the human brain. The focus of this work is on the design of an architecture to support HMAX, a widely accepted model of the human visual pathway. The computationally intensive nature of HMAX and wide applicability in real-time visual analysis application makes the design of hardware accelerators a key necessity. In this work, we propose a configurable accelerator mapped efficiently on a FPGA to realize real-time feature extraction for vision-based classification algorithms. Our innovations include the efficient mapping of the proposed architecture on the FPGA as well as the design of an efficient memory structure. Our evaluation shows that the proposed approach is significantly faster than other contemporary solutions on different platforms.

7.3.1.3 Power-Efficient Error-Resiliency for H.264/AVC Context-Adaptive Variable Length Coding [p. 697]
M Shafique, B Zatt, S Rehman, F Kriebel and J Henkel

Technology scaling has led to unreliable computing hardware due to high susceptibility against soft errors. In this paper, we propose an error-resilient architecture for Context-Adaptive Variable Length Coding (CAVLC) in H.264/AVC. Due to its context-adaptive nature and intricate control flow CAVLC is very sensitive to soft errors. An error during the CAVLC process (especially during the context adaptation or in VLC tables) may result in severe mismatch between encoder and decoder. The primary goal in our error-resilient CAVLC architecture is to protect codeword/codelength tables and context adaptation in reliable yet power efficient manner. For reducing the power overhead, the tables are partitioned in various sub-tables each protected with variable-sized parity. Moreover, for further power reduction, our approach incorporates state-retentive power-gating of different sub-tables at run time depending upon the statistical distribution of syntax elements. Compared to the unprotected case, our scheme provides a video quality improvement of 18dB (averaged over various fault injection cases and video sequences) at the cost of 35% area overhead and 45% performance overhead due to the error-detection logic. However, partitioned sub-tables increase the potential for power-gating, thus bring a leakage energy saving of 58%. Compared to state-of-the-art table protection, our scheme provides 2x reduced area and performance overhead. For functional verification and area comparison, the architecture is prototyped on a Xilinx Virtex-5 FPGA, though not limited to it. For the soft errors experiments, evaluation of error-resiliency and power efficiency, we have developed a fault injection and simulation setup.

7.3.1.4 Towards Accurate Hardware Stereo Correspondence: A Real-Time FPGA Implementation of a Segmentation-Based Adaptive Support Weight Algorithm [p. 703]
C Ttofis and T Theocharides

Disparity estimation in stereoscopic vision is a vital step for the extraction of depth information from stereo images. This paper presents the hardware implementation of a disparity estimation system that enables good performance in both accuracy and speed. The architecture implements an adaptive support weight stereo correspondence algorithm, which integrates information obtained from image segmentation, in an attempt to increase the robustness of the matching process. The proposed system integrates optimization techniques that make the algorithm hardware-friendly and suitable for embedded vision systems. A prototype of the architecture was implemented on an FPGA, achieving 30 fps for 640x480 image sizes. The quality of the disparity maps generated by the proposed system is also better than other existing hardware implementations featuring fixed support local correspondence methods.
Nanoelectronic Devices

Financial engineering is a very active research field as a result of the growth of the derivative markets and the complexity of the mathematical models utilized in pricing the numerous financial products. In this paper, we present an FPGA-based parallel processor optimized for solving the Black-Scholes partial derivative equation utilized in option pricing which employs the two most widely used finite difference schemes: Crank-Nicholson and explicit differences. As our measurements demonstrate, the presented architecture is expandable and the speedup triggered is increased almost linearly with the available silicon resources. Although the processor is optimized for this specific application, it is highly programmable and thus it can significantly accelerate all applications that use finite differences computations. Performance measurements show that our FPGA prototype triggers a 5x speedup when compared with a 2GHz dual-core Intel CPU (Core2Duo). Moreover, for the explicit scheme, our FPGA processor provides an 8x speedup over the same Intel processor.

Keywords: Option pricing, finite differences, FPGA

7.4: Nanoelectronic Devices

Moderators: S Garg, Toronto U, CA; C Nicopoulos, Cyprus U, CY

A SAT-based Fitness Function for Evolutionary Optimization of Polymorphic Circuits [p. 715]
L Sekanina and Z Vasicek

Multifunctional (or polymorphic) gates have been utilized as building blocks for multifunctional circuits that are capable of performing various logic functions under different settings of control signals. In order to effectively synthesize polymorphic circuits, several methods have been developed in the recent years. Unfortunately, the methods are applicable for small circuits only. In this paper, we propose a SAT-based functional equivalence checking algorithm to eliminate the fitness evaluation time which is the most critical overhead for genetic programming-based design and optimization of complex polymorphic circuits. The proposed approach has led to a 20%-40% reduction in gate count with respect to the solutions created using the polymorphic multiplexing.

Mach-Zehnder Interferometer Based Design of All Optical Reversible Binary Adder [p. 721]
S Kotiyal, H Thapliyal and N Ranganathan

In recent years reversible logic has emerged as a promising computing model for applications in dissipation less optical computing, low power CMOS, quantum computing, etc. In reversible circuits there exist a one-to-one mapping between the inputs and the outputs resulting in no loss of information. Researchers have implemented reversible logic gates in optical computing domain as it can provide high speed and low energy requirement along with easy fabrication at the chip level [1]. The all optical implementation of reversible gates are based on semiconductor optical amplifier (SOA) based Mach-Zehnder interferometer (MZI) due to its significant advantages such as high speed, low power, fast switching time and ease in fabrication. In this work we present the all optical implementation of an n bit reversible ripple carry adder for the first time in literature. The all optical reversible adder design is based on two new optical reversible gates referred as optical reversible gate I (ORG-I) and optical reversible gate II (ORG-II) and the existing all optical Feynman gate. The two new reversible gates ORG-I and ORGII are proposed as they can implement a reversible adder with reduced optical cost which is the measure of number of MZIs switches and the propagation delay, and with zero overhead in terms of number of ancilla inputs and the garbage outputs. The proposed all optical reversible adder design based on the ORG-I and ORG-II reversible gates are compared and shown to be better than the other existing designs of reversible adder proposed in non-optical domain in terms of number of MZIs, delay, number of ancilla inputs and the garbage outputs. The proposed all optical reversible ripple carry adder will be a key component of an all optical reversible ALU that can be applied in a wide variety of optical signal processing applications.

Weighted Area Technique for Electromechanically Enabled Logic Computation with Cantilever-Based NEMS Switches [p. 727]
S Patil, M-W Jang, C-L Chen, D Lee, Z Ye, W E Partlo III, D J Lilja, S A Campbell and T Cui

Nanoelectromechanical systems (NEMS) is an emerging nanoscale technology that combines mechanical and electrical effects in devices. A variety of NEMS-based devices have been proposed for integrated chip designs. Amongst them are near-ideal digital switches. The electromechanical principles that are the basis of these switches impart the capability of extremely low power switching characteristics to digital circuits. NEMS switching devices have been mostly used as simple switches to provide digital operation, however, we observe that their unique operation can be used to accomplish logic functions directly. In this paper, we propose a novel technique called “weighted area logic” to design logic circuits with NEMS-based switches. The technique takes advantage of the unique structural configurations possible with the NEMS devices to convert the digital switch from a simple ON-OFF switch to a logical switch. This transformation not only reduces the delay of complex logic units, but also decreases the power and area of the implementation further. To demonstrate this, we show the new
designs of the logic functions of NAND, XOR and a three input function $Y = A + B : C$, and compose them into a 32-bit adder. Through simulation, we quantify the power, delay and area advantages of using the weighted area logic technique over a standard CMOS-like design technique applied to NEMS.

7.5: High Level and Statistical Design of Mixed-Signal Systems

Moderators: C Dehollain, EPF Lausanne, CH; D Morche, CEA-Leti, FR

L Wang, T J Kazmierski, B M Al-Hashimi, M Aloufi and J Wenninger

In an energy harvester powered wireless sensor node, the energy harvester is often the only energy source, therefore it is crucial to configure the microcontroller and the sensor node so that the harvested energy is used efficiently. This paper presents a response surface model (RSM) based design space exploration and optimisation of a complete wireless sensor node system. In our work the power consumption models of the microcontroller and the sensor node are defined based on their digital operations so that the parameters of the digital algorithms can be optimised to achieve the best energy efficiency. In the proposed technique, SystemC-A is used to model the system's analogue components as well as the digital control algorithms implemented in the microcontroller and the sensor node. A series of simulations are carried out and a response surface model is constructed from the simulation results. The RSM is then optimised using MATLAB's optimisation toolbox and the results show that the optimised system configuration can double the total number of wireless transmissions with fixed amount of harvested energy. The great improvement in the system performance validates the efficiency of our technique.

A Leveque, F Pecheux, M-M Louerat, H Aboushady, F Cenni, S Scotti, A Massouri and L Clavier

The paper presents the principles, techniques and tools for the efficient modeling and simulation, at the component level, of an heterogeneous system composed of Wireless Sensor Network nodes that exhibits complex multi-discipline feedback loops that are likely to be found in many state-of-the-art applications such as cyber-physical systems. A Precollision Mitigation Braking System (PMBS) is used as a pragmatic case study to validate the whole approach. The component models presented (60 GHz communication channel, QPSK RF transceiver, CMOS video sensor, digital microcontroller, simplified car kinetic engine) are written in SystemC and its analog Mixed-Signal extensions, SystemC-AMS, and belong to five distinct yet highly interwoven disciplines: newtonian mechanics, opto-electronics, analog RF, digital and embedded software. The paper clearly exhibits the complex multi-discipline feedback loop of this automotive application and the related model composability issues. Using the opto-electrical stimulus and the received RF inter-vehicle data, a car is able to exploit its environmental data to autonomously adjust its own velocity. This adjustment impacts the physical environment that in turns modifies the RF communication conditions. Results show that this holistic first-order virtual prototype can be advantageously used to jointly develop the final embedded software and to refine any of its hardware component part.

E Maricau, D De Jonghe and G Gielen

The paper discusses a technique to perform efficient circuit reliability analysis of large analog and mixed-signal systems. The proposed method includes the impact of both process variations and transistor aging effects. The complexity of large systems is dealt with by partitioning the system into manageable subblocks that are modeled separately. These models are then evaluated to obtain the system specifications. However, highly expensive reliability simulations, combined with nonlinear output behavior and the high dimensionality of the problem is still a very challenging task. Therefore the use of fast function extraction symbolic regression (FFX) is proposed. This allows to capture the high-dimensional nonlinear problem with good accuracy. Also, an active learning sample selection algorithm is introduced to minimize the amount of expensive aging simulations. The algorithm trades of space exploration with function nonlinearity detection and model uncertainty reduction to select optimal model training samples. The simulation method is demonstrated on a 6 bit Flash ADC, designed in a 32nm CMOS technology. Experimental results show a speedup of 360x over existing aging simulators to evaluate 100 Monte-Carlo samples with good accuracy.

B Liu, J Messaoudi and G Gielen

Yield estimation for analog integrated circuits remains a time-consuming operation in variation-aware sizing. State-of-the-art statistical methods such as ranking-integrated Quasi-Monte-Carlo (QMC), suffer from performance degradation if the number of effective variables is
large (as typically is the case for realistic analog circuits). To address this problem, a new method, called AYLeSS, is proposed to estimate the yield of analog circuits by introducing Latin Supercube Sampling (LSS) technique from the computational statistics field. Firstly, a partitioning method is proposed for analog circuits, whose purpose is to appropriately partition the process variation variables into low-dimensional sub-groups fitting for LSS sampling. Then, randomized QMC is used in each sub-group. In addition, the way to randomize the run order of samples in Latin Hypercube Sampling (LHS) is used for the QMC sub-groups. AYLeSS is tested on 4 designs of 2 example circuits in 0.35μm and 90nm technologies with yield from about 50% to 90%. Experimental results show that AYLeSS has approximately a 2 times speed enhancement compared with the best state-of-the-art method.

Keywords Yield estimation, analog circuits, Latin Supercube Sampling (LSS)

This contribution presents a major improvement for our analog synthesis framework with an explorative characteristic. The presented approach in principle allows the synthesis of a wide range of circuits, without the limitation to specific circuit classes. Defined by a specification of up to 15 different performances, a fully sized, transistor level circuit is synthesized for a provided process technology. The presented work reduces the needed computational effort and thus drastically reduces the synthesis time, while adding new abstraction into the framework to provide an even wider range of synthesized circuits - demonstrated in experimental results.

7.6: Advances in Dataflow Modeling and Analysis

Moderators: C Haubelt, Rostock U, DE; L S Indrusiak, York U, UK

This paper presents a cost-effective and formal approach to model and analyze streaming applications on multiprocessor systems-on-chip (MPSoCs). This approach enables to address time requirements, mapping of applications on MPSoCs and system behavior correctness by using abstract clocks of synchronous languages. Compared to usual prototyping and simulation techniques, it is very fast and favors correctness-by-construction. No coding is needed to run and analyze a system, which avoids tedious debugging efforts. It is an ideal complement to existing techniques to deal with large system design spaces.

Synchronous dataflow graphs (SDFGs) are used extensively to model streaming applications. An SDFG can be extended with scheduling decisions, allowing SDFG analysis to obtain properties like throughput or buffer sizes for the scheduled graphs. Analysis times depend strongly on the size of the SDFG. SDFGs can be statically scheduled using static-order schedules. The only generally applicable technique to model a static-order schedule in an SDFG is to convert it to a homogeneous SDFG (HSDFG). This conversion may lead to an exponential increase in the size of the graph and to sub-optimal analysis results (e.g., for buffer sizes in multi-processors). We present a technique to model periodic static-order schedules directly in an SDFG. Experiments show that our technique produces more compact graphs compared to the technique that relies on a conversion to an HSDFG. This results in reduced analysis times for performance properties and tighter resource requirements.

System-level design space exploration (DSE), which is performed early in the design process, is of eminent importance to the design of complex multi-processor embedded system architectures. During system-level DSE, system parameters like, e.g., the number and type of processors, the type and size of memories, or the mapping of application tasks to architectural resources, are considered. Simulation-based DSE, in which different design instances are evaluated using system-level simulations, typically are computationally costly. Even using high-
level simulations and efficient exploration algorithms, the simulation time to evaluate design points forms a real bottleneck in such DSE. Therefore, the vast design space that needs to be searched requires effective design space pruning techniques. This paper presents a technique to reduce the number of simulations needed during system-level DSE. More specifically, we propose an iterative design space pruning methodology based on static throughput analysis of different application mappings. By interleaving these analytical throughput estimations with simulations, our hybrid approach can significantly reduce the number of simulations that are needed during the process of DSE.

7.7: Test and Repair of New Technologies
Moderators: J Tyszer, TU Poznan, PL; H-J Wunderlich, Stuttgart U, DE

[166] Test Pin Count Reduction for NoC-based Test Delivery in Multicore SOCs [p. 787]
M Richter and K Chakrabarty

We present the first pin-count-aware optimization approach for test data delivery over a network-on-chip (NoC). By co-optimizing core test scheduling and pin assignment to access points, the limited I/O resources provided by automated test equipment (ATE) can be used more effectively. This approach allows us to lower test cost by reducing test time for a given pin budget, or by reducing the number of test pins without impacting test time. To further improve resource utilization, we consider the use of MISRs for compacting the test responses of embedded cores. Experimental results for ITC'02 test benchmarks demonstrate that pin-count-aware co-optimization leads to shorter test times for a given pin-count budget and fewer pins for a given test-time budget. The results also highlight the advantages of the proposed use of output compaction.

[167] On Effective TSV Repair for 3D-Stacked ICs [p. 793]
L Jiang, Q Xu and B Eklow

3D-stacked ICs that employ through-silicon vias (TSVs) to connect multiple dies vertically have gained wide-spread interest in the semiconductor industry. In order to be commercially viable, the assembly yield for 3D-stacked ICs must be as high as possible, requiring TSVs to be repairable. Existing techniques typically assume TSV faults to be uniformly distributed and use neighboring TSVs to repair faulty ones, if any. In practice, however, clustered TSV faults are quite common due to the fact that the TSV bonding quality depends on surface roughness and cleanliness of silicon dies, rendering prior TSV redundancy solutions less effective. To resolve this problem, we present a novel TSV repair framework, including a hardware architecture that enables faulty TSVs to be repaired by redundant TSVs that are farther apart, and the corresponding repair algorithm. By doing so, the manufacturing yield for 3D-stacked ICs can be dramatically improved, as demonstrated in our experimental results.

N Z Haron and S Hamdioui

Resistive random access memory (RRAM) is one of the universal memory candidates for computer systems. Although RRAM promises many attractive advantages (e.g., huge data storage, smaller form-factor, lower power consumption, non-volatility, etc.), there are many open issues that still need to be solved, especially those related to its quality and reliability. For instance, open defects may cause RRAM cell to enter an undefined state (i.e., somewhere between logic 0 and 1), making it hard to detect during manufacturing test. As a consequence, this may lead to test escapes (quality issue) and field failures (reliability issue). This paper shows - based on defect and circuit simulation - how testing RRAM is different from testing conventional random access memories and how march test cannot guarantee higher defect coverage. The paper then motivates the need of development of special Design-for-Testability (DfT). A concept of a new DfT is then proposed. The concept is further exploited and mapped into two different DfT circuits (i) Short Write Time and (ii) Low Write Voltage. Both DfT schemes are implemented and simulated; the simulation results show that defects causing the RRAM cell to enter an undefined state are easily detected. Keywords - quality, reliability, memory defect, Design-for-Testability, memristor.

7.8: HOT TOPIC - New Directions in Timing Modeling and Analysis of Automotive Software
Moderator: W Mueller, U Paderborn, DE

M-A Peraldi-Frati, H Blom, D Karlsson and S Kuntz

In the automotive industry, the Automotive Open System Architecture AUTOSAR is established as a de-facto standard and is applied in a steadily increasing number of development projects. In addition, AUTOSAR attracted the attention of other non-automotive industries, like railway, agriculture and construction machines, power generation and marine technology. The first versions of the standard successfully...
achieved the objective of integrating in a common framework various components from different suppliers and ensuring their interfaces interoperability. In actual and future versions of the standard, the objective becomes even more ambitious as it considers behavioral and timing characteristics of these components. Therefore, this paper presents the current status of AUTOSAR Release 4.0 concerning the behavioral modeling and timing characterization of components and opens several research and development directions for future extensions of the standard.

Keywords: Timing Modeling; Timing Analysis, AUTOSAR; EAST-ADL, Multiform Time, Probabilistic Timing

Modeling and analysis of timing information are essential to the design of real-time systems. In this domain, research related to probabilistic analysis is motivated by the desire to refine results obtained using worst-case analysis for systems in which the worst-case scenario is not the only relevant one, such as soft real-time systems. This paper presents an overview of the existing solutions for probabilistic timing analysis, focusing on challenges they have to face. We discuss in particular two new trends toward Probabilistic Real-Time Calculus and Typical-Case Analysis which rise to some of these challenges.

Index Terms - Real-time systems, Stochastic analysis, Probabilistic Real-Time Calculus, Typical-Case Analysis.

IP3: Interactive Presentations

QBF-Based Boolean Function Bi-Decomposition [p. 816]
H Chen, M Janota and J Marques-Silva

Boolean function bi-decomposition is ubiquitous in logic synthesis. It entails the decomposition of a Boolean function using two-input simple logic gates. Existing solutions for bi-decomposition are often based on BDDs and more recently, on Boolean Satisfiability. In addition, the partition of the input set of variables is either assumed, or heuristic solutions are considered for finding good partitions. In contrast to earlier work, this paper proposes the use of Quantified Boolean Formulas (QBF) for computing bi-decompositions. These bi-decompositions are optimal in terms of the achieved quality of the input set of variables. Experimental results, obtained on representative benchmarks, demonstrate clear improvements in the quality of computed decompositions, but also the practical feasibility of QBF-based bi-decomposition.

Automatic Transition Between Structural System Views in a Safety Relevant Embedded Systems Development Process [p. 820]
C Ellen, C Etzion and M Oertel

It is mandatory to design safety relevant embedded systems in multiple structural system views. A typical example is the usage of a functional and technical system representation. A transition between these system views not only comprises the allocation of components but also copes with multiple design aspects and constraints that need to be transferred to the target perspective. Optimization goals regarding arbitrary design artifacts complicate this problem. In this paper we present a novel comprehensive approach integrating common allocation techniques together with a partial design generation in a system wide process to optimize complex system view transitions. We demonstrate our approach using the CESAR design methodology. The original system models and requirements are used as input for our procedure and the results are directly applied to the same models.

Towards New Applications of Multi-Function Logic: Image Multi-Filtering [p. 824]
L Sekanina and V Salajka

Multifunctional (or polymorphic) gates are capable of performing two or more logic functions according to the setting of control signals. They can be considered as building blocks for new and cheap reconfigurable chips. In this paper, we utilized multifunctional components that can be implemented using multifunctional gates as building blocks of image filters. We applied genetic programming to evolve image filters performing different filtering tasks under different settings of control signals. Evolved solutions exhibit a significant reduction in utilized operations and interconnects w.r.t. the multiplexing of conventional solutions.

Memory-Map Selection for Firm Real-Time SDRAM Controllers [p. 828]
S Goossens, T Kouters, B Akesson and K Goossens

A modern real-time embedded system must support multiple concurrently running applications. To reduce costs, critical SoC components like SDRAM memories are often shared between applications with a variety of firm real-time requirements. To guarantee that the system works as intended, the memory controller must be configured such that all the real-time requirements of all sharing applications are satisfied. The attainable worst-case bandwidth, latency, and power of the memory depend largely on memory map configuration. Sharing SDRAM amongst multiple applications is challenging, since their requirements might call for different memory maps. This paper presents an exploration of the
memory-map design space. Two contributions improve the memory-map selection procedure. The first contribution reduces the minimum access granularity by interleaving requests over a configurable number of banks instead of all banks. This technique is beneficial for worst-case performance in terms of bandwidth, latency and power. As a second contribution, we present a methodology to derive a memory-map configuration, i.e. the access granularity and number of interleaved banks, from a specification of the real-time application requirements and an overall memory power budget.

Y Liang, Z Cui, S Zhao, K Rupnow, Y Zhang, D L Jones and D Chen

Real-time 3D sound localization is an important technology for various applications such as camera steering systems, robotics audition, and gunshot direction. 3D sound localization adds a new dimension, but also significantly increases the computational requirements. Real-time 3D sound localization continuously processes large volumes of data for each possible 3D direction and acoustic frequency range. Such highly demanding compute requirements outpace current CPU compute abilities. This paper develops a real-time implementation of 3D sound localization on Graphical Processing Units (GPUs). Massively parallel GPU architectures are shown to be well suited for 3D sound localization. We optimize various aspects of GPU implementation, such as number of threads per thread block, register allocation per thread, and memory data layout for performance improvement. Experiments indicate that our GPU implementation achieves 501X and 130X speedup compared to a single-thread and a multi-thread CPU implementation respectively, thus enabling real-time operation of 3D sound localization.

A Xhakoni, D San Segundo Bello and G Gielen

This paper introduces a 3D-integrated image sensor with high dynamic range, high frame rate and high resolution capabilities. A robust algorithm for dynamic range extension with low sensitivity to circuit non-idealities and based on multiple exposures is presented. The impact of the TSV diameter over the dynamic range and frame rate performance is studied allowing the choice of the best 3D technology for the required performance.

Keywords-component; 3D integration, CMOS image sensor, high frame rate, high dynamic range

M J Dousti and M Pedram

Quantum computers are exponentially faster than their classical counterparts in terms of solving some specific, but important problems. The biggest challenge in realizing a quantum computing system is the environmental noise. One way to decrease the effect of noise (and hence, reduce the overhead of building fault tolerant quantum circuits) is to reduce the latency of the quantum circuit that runs on a quantum circuit. In this paper, a novel algorithm is presented for scheduling, placement, and routing of a quantum algorithm, which is to be realized on a target quantum circuit fabric technology. This algorithm, and the accompanying software tool, advances state-of-the-art in quantum CAD methodologies and methods while considering key characteristics and constraints of the ion-trap quantum circuit fabric. Experimental results show that the presented tool improves results of the previous tool by about 41%.

Keywords- quantum computing; scheduling; routing; placement; ion-trap technology; CAD tool

C Zhang, V F Pavlidis and G De Micheli

Power grid analysis is a challenging problem for modern integrated circuits. For 3-D systems fabricated using stacked tiers with TSVs, traditional power grid analysis methods for planar (2-D) circuits do not demonstrate the same performance. An efficient IR drop analysis method for 3-D large-scale circuits, called 3-D voltage propagation method, is proposed in this paper. This method is compared with another widely used power grid analysis method, with preconditioned conjugated gradients. Simulation results demonstrate that the proposed method is more efficient for the IR drop analysis of large size 3-D power grids. Speedups between 10x to 20x over the preconditioned conjugated gradients method are shown.

Keywords - 3-D integrated circuits; Through-silicon via; Power grid analysis

S A Nazin, D Morche and A Reinhardt

This paper is devoted to the yield optimization of the radio-frequency (RF) front-end of wireless receiver. The yield together with the circuit performances are often sensitive to the choice of parameters of its components and blocks, and can be improved at circuit design level. However, it is better evaluated when considering the whole receiver at system level. For this purpose, we first use a design of experiment (DoE) technique to generate meta-models of the building blocks. Then, we apply a version of the stochastic gradient method to find a good
approximation of the optimum for the circuit yield.

[180] Parallel Statistical Analysis of Analog Circuits by GPU-accelerated Graph-based Approach [p. 852]

X-X Liu, S X-D Tan and H Wang

In this paper, we propose a new parallel statistical analysis method for large analog circuits using determinant decision diagram (DDD) based graph technique based on GPU platforms. DDD-based symbolic analysis technique enables exact symbolic analysis of very large analog circuits. But we show that DDD-based graph analysis is very amenable for massively threaded based parallel computing based on GPU platforms. We design novel data structures to represent the DDD graphs in the GPUs to enable fast memory access of massive parallel threads for computing the numerical values of DDD graphs. The new method is inspired by inherent data parallelism and simple data independence in the DDD-based numerical evaluation process. Experimental results show that the new evaluation algorithm can achieve about one to two order of magnitudes speedup over the serial CPU based evaluations and 2-3 times speedup over numerical SPICE-based simulation method on some large analog circuits.


R Rudolf, P Taatizadeh, R Wilcock and P Wilson

A novel approach is proposed for analogue circuits that identifies which devices should be replaced with configurable analogue transistors (CATs) to maximise post fabrication yield. Both performance sensitivity and adjustment independence are considered when identifying these critical devices, giving a combined weighted sensitivity. The results from an operational amplifier case study are presented where it is demonstrated that variation in key circuit performances can be reduced by an average of 78.8% with the use of only three CATs. These results confirm that the proposed critical device selection method with optimal performance driven CAT sizing can lead to significant improvement in overall performance and yield.

Keywords-configurable analogue transistor; optimal sizing; device variability; sensitivity analysis, post fabrication calibration


J Zimmermann, O Bringmann and W Rosenstiel

Synchronous dataflow (SDF) models are gaining increased attention in designing software-intensive embedded systems. Especially in the signal processing and multimedia domain, dataflow-oriented models of computation are commonly used by designers reflecting the regular structure of algorithms and providing an intuitive way to specify both sequential and concurrent system functionality. Furthermore, dataflow-oriented models are qualified for capturing dynamic behavior due to data-dependent execution. In this work, we extend those data-dependent dataflow models to include dynamic power management (DPM) aspects of a target platform while still meeting hard timing requirements. We capture different system states in a multi-domain scenario approach and develop a state space based on this SDF representation for system analysis and optimization. By traversing the state space of the power-aware scenario modeling we derive a power management configuration with minimized energy dissipation depending on dynamic system behavior.


A Das, U Kocabas, A-R Sadeghi and I Verbauwhede

Globalization of the semiconductor industry increases the vulnerability of integrated circuits. This particularly becomes a major concern for cryptographic IP blocks integrated on a System-on-Chip (SoC). The trustworthiness of these cryptographic blocks can be ensured with a secure test strategy. Presently, the IEEE 1500 Test Wrapper has emerged as the test standard for industrial SoCs. Additionally a secure activation mechanism has been proposed to this standard in order to restrict access to the testing interface to eligible testers by using a cryptographic authentication mechanism. This access mechanism is necessary in order not to provide any side-channels which may leak secret information for attackers. However, this approach requires the authentication mechanism to be implemented in hardware incurring an area overhead, and the authentication secrets to be securely stored in non-volatile memory (NVM), which may be susceptible to side-channel attacks. In this work, we enhance the secure test wrapper allowing testing of multiple IP blocks using a PUF-based authentication mechanism which overcomes the necessity of secure NVM and reduces the implementation overhead.

Keywords- Secure Test Wrapper; Scan Chains; SoC Testing; Physically Unclonable Functions (PUF)

8.1: HOT TOPIC - Robustness Challenges in Automotive (Special Day E-Mobility)

Moderator: J Lau, Infineon, DE

[184] Complexity, Quality and Robustness - The Challenges of Tomorrow's Automotive Electronics [p. 870]

U Abelein, H Lochner, D Hahn and S Straube
Developing a state-of-the-art premium car means implementing one of the most complex electronic systems mankind is using in daily life. About 100 ECU's with more than 7,000 semiconductor components realize safety, comfort and powertrain functions. These numbers will increase drastically when going the step from the conventional vehicle to the e-car, where a lot more functions have to be realized just by electronics. Finally the functionality has to be guaranteed in one of the harshest environments where electronics are used with a target of 0 ppm concerning the subcomponents for the 15 years lifetime of the car. We give an overview of the challenges on the way to reach this target. The task of getting to a high level of robustness and quality within short maturing periods of new technologies and semiconductor products are discussed. Using state-of-the-art semiconductor process technologies for devices in a car is necessary to fulfill today’s performance requirements and even more future requirements with respect to the e-car. But it leads to a mission which seems to be a paradoxon: combining more robustness of the complete system and quality of its subcomponents with less mature technologies. A way out of this dilemma can only be found by reviewing carefully today's qualification and validation processes and understand their strengths, weaknesses and capabilities. This must be the starting point for an evolution to a qualification strategy which is suitable for this fundamentally changed situation. Therefore the limits of today's qualification methods will be discussed as well as some suggestions for future strategies will be made to bring complexity, quality and robustness in an early phase of product lifetime together. The roles of the parties in the supply chain shall be highlighted in these strategies as well.

Keywords: automotive, semiconductor, robustness, qualification

Automotive power micro-electronic devices in the past were low pin-count, low complexity devices. Robustness could be assessed by stressing the few operating conditions and by manual analysis of the simple analog circuitry. Nowadays complexity of Automotive Smart Power Devices is driven by the demands for energy efficiency and safety, which adds the need for additional monitoring circuitry, redundancy, power-modes, leading even to complex System-on-chips with embedded uC cores, embedded memory, sensors and other elements. Assessing the application robustness of this type of microelectronic devices goes hand-in-hand with exploring their verification space inside and to certain extent outside of the specification. While there are well established methods for standard functional verification, methods for application oriented robust verification are not yet available. In this paper we present promising directions and first results, to explore and assess device robustness through various pre- and post-Si verification and design exploration strategies, focusing on metamodeling, constrained-random verification and hardware-in-the-loop experiments, for exploration of the operating space.

Keywords: component; Automotive Smart Power IC, Robustness, Metamodeling, Constrained-Random-Verification

8.2: PANEL - EDA for Trailing Edge Technologies

Moderator: P Rolandi STMicroelectronics Italy

Over the last decade, the semiconductor industry has advanced CMOS technology from 90 to 22/20 nanometers, and the EDA industry has developed a great deal of tools, methodologies, and flows to help “gigascale” design, implementation and verification, at these “leading edge” technology nodes. However, in 2010 approximately 75% of design starts used 130 nanometers or greater CMOS technologies [1], and 25% of wafers were fabricated using these “trailing edge” technologies [2]. There are possibly more designers working at 130 nanometers and above than at 90 nanometers and below, and there is certainly much more to electronics than just digital CMOS and microprocessors, and in order for the electronic industry to continue delivering on promises, “More than Moore” is needed, besides “More of Moore”. What is EDA doing - or what should EDA do - in order to help design implementation and verification at trailing edge technologies?

8.3: Innovative Reliable Systems and Applications

Moderators: J Ayala, Madrid Complutense U, ES; M D Santambrogio, Politecnico di Milano, IT

Checkpoint and Recovery (CR) allows computer systems to operate correctly even when compromised by transient faults. While many software systems and hardware systems for CR do exist, they are usually either too large, require major modifications to the software, too slow, or require extensive modifications to the caching schemes. In this paper, we propose a novel error-recovery management scheme, which is based upon re-engineering the instruction set. We take the native instruction set of the processor and enhance the microinstructions with additional micro-operations which enable checkpointing. The recovery mechanism is implemented by three custom instructions, which recover the registers which were changed, the data memory values which were changed and the special registers (PC, status registers etc.) which were changed.

Keywords: Hardware/Software Checkpoint and Recovery Scheme for Embedded Processors
Our checkpointing storage is changed according to the benchmark executed. Results show that our method degrades performance by just 1.45% under fault free conditions, and incurs area overhead of 45% on average and 79% in the worst case. The recovery takes just 62 clock cycles (worst case) in the examples which we examined.

C Zambelli, M Indaco, M Fabiano, S Di Carlo, P Prinetto, P Olivo and D Bertozzi

In spite of the mature cell structure, the memory controller architecture of Multi-level cell (MLC) NAND Flash memories is evolving fast in an attempt to improve the uncorrected/miscorrected bit error rate (UBER) and to provide a more flexible usage model where the performance-reliability trade-off point can be adjusted at runtime. However, optimization techniques in the memory controller architecture cannot avoid a strict trade-off between UBER and read throughput. In this paper, we show that co-optimizing ECC architecture configuration in the memory controller with program algorithm selection at the technology layer, a more flexible memory sub-system arises, which is capable of unprecedented trade-offs points between performance and reliability.

A Resilient Architecture for Low Latency Communication in Shared-L1 Processor Clusters [p. 887]
M R Kakoee, I Loi and L Benini

A reliable and variation-tolerant architecture for shared-L1 processor clusters is proposed. The architecture uses a single-cycle mesh of tree as the interconnection network between processors and a unified Tightly Coupled Data Memory (TCDM). The proposed technique is able to compensate the effect of process variation on processor to memory paths. By adding one stage of controllable pipeline on the processor to memory paths we are able to switch between two modes: with and without pipeline. If there is no variation, the processor to memory path is fully combination and we have single-cycle read and write operations. If the variation occurs, the controllable pipeline is switched to pipeline mode and by increasing the latency of the read/write operation we mitigate the effect of the variations. We also propose a configuration-time approach to conditionally add the extra pipeline state based on detection of timing-critical paths. Experimental results show that our speed adaptation approach is able to compensate up-to 90% degradation in the request path with less than 1% hardware overhead for a shared-L1 CMP with 16 processors and 32 memory banks. We show that even if variation occurs on all processor to memory paths, our approach can mitigate it with an average overhead of 20% on the application's runtime.

Performance-Reliability Tradeoff Analysis for Multithreaded Applications [p. 893]
I Oz, H R Topcuoglu, M Kandemir and O Tosun

Modern architectures become more susceptible to transient errors with the scale down of circuits. This makes reliability an increasingly critical concern in computer systems. In general, there is a tradeoff between system reliability and performance of multithreaded applications running on multicore architectures. In this paper, we conduct a performance-reliability analysis for different parallel versions of three data-intensive applications including FFT, Jacobi Kernel, and Water Simulation. We measure the performance of these programs by counting execution clock cycles, while the system reliability is measured by Thread Vulnerability Factor (TVF) which is a recently-proposed metric. TVF measures the vulnerability of a thread to hardware faults at a high level. We carry out experiments by executing parallel implementations on multicore architectures and collect data about the performance and vulnerability. Our experimental evaluation indicates that the choice is clear for FFT application and Jacobi Kernel. Transpose algorithm for FFT application results in less than 5% performance loss while the vulnerability increases by 20% compared to binary-exchange algorithm. Unrolled Jacobi code reduces execution time up to 50% with no significant change on vulnerability values. However, the tradeoff is more interesting for Water Simulation where nsquared version reduces the vulnerability values significantly by worsening the performance with similar rates compared to faster but more vulnerable spatial version.

Index Terms - Multi-Core Architectures and Support, Reliable Parallel and Distributed Algorithms

8.4: Advances in Formal SoC Verification

Efficient Groebner Basis Reductions for Formal Verification of Galois Field Multipliers [p. 899]
J Lv, P Kalla and F Enescu

Galois field arithmetic finds application in many areas, such as cryptography, error correction codes, signal processing, etc. Multiplication lies at the core of most Galois field computations. This paper addresses the problem of formal verification of hardware implementations of (modulo) multipliers over Galois fields of the type $F_{2^k}$ using a computer-algebra/algebraic-geometry based approach. The multiplier circuit is modeled as a polynomial system in $F_{2^k} [x_1, x_2, \ldots, x_d]$ and the verification problem is formulated as a membership test in a corresponding (radical) ideal. This requires the computation of a Gröbner basis, which can be computationally intensive. To overcome this limitation, we analyze the circuit topology and derive a term order to represent the polynomials. Subsequently, using the theory of Gröbner bases over Galois...
fields, we prove that this term order renders the set of polynomials itself a Gröbner basis of this ideal - thus significantly improving verification. Using our approach, we can verify the correctness of, and detect bugs in, up to 163-bit circuits in $F_{2^{163}}$; whereas contemporary approaches are infeasible.

[192] Scalable Progress Verification in Credit-Based Flow-Control Systems [p. 905]
S Ray and R K Brayton

Formal verification of liveness properties of practical communication fabrics are generally intractable with present day verification tools. We focus on a particular type of liveness called "progress" which is a form of deadlock freedom. An end-to-end progress property is broken down into localized safety assertions, which are more easily provable, and lead to a formal proof of progress. Our target systems are credit-based flow-control networks. We present case studies of this type and experimental results of progress verification of large networks using a bit-level formal verifier.

S Mitra, A Banerjee and P Dasgupta

Bug-fixing in deeply embedded portions of the logic is typically accompanied by the post-facto addition to new assertions which cover the bug scenario. Formally verifying properties defined over such deeply embedded portions of the logic is challenging because formal methods do not scale to the size of the entire logic, and verifying the property on the embedded logic in isolation typically throws up a large number of counterexamples, many of which are spurious because the scenarios they depict are not possible in the entire logic. In this paper we introduce the notion of ranking the counterexamples so that only the most likely counterexamples are presented to the designer. Our ranking is based on assume properties mined from simulation traces of the entire logic. We define a metric to compute a belief for each assume property that is mined, and rank counterexamples based on their conflicts with the mined assume properties. Experimental results demonstrate an amazing correlation between the real counterexamples (if they exist) and the proposed ranking metric, thereby establishing the proposed method as a very promising verification approach.

8.5: Variability and Delay

Moderators: S Sapatnekar, Minnesota U, US; J Cortadella, UP Catalunya, ES

[194] Transistor-Level Gate Model Based Statistical Timing Analysis Considering Correlations [p. 917]
Q Tang, A Zjajo, M Berkelaar and N van der Meijns

To increase the accuracy of static timing analysis, the traditional nonlinear delay models (NLDMs) are increasingly replaced by the more physical current source models (CSMs). However, the extension of CSMs into statistical models for statistical timing analysis is not easy. In this paper, we propose a novel correlation-preserving statistical timing analysis method based on transistor-level gate models. The correlations among signals and between process variations are fully accounted for. The accuracy and efficiency are obtained from statistical transistor-level gate models, evaluated using a smart Random Differential Equation (RDE)-based solver. The variational waveforms are available, allowing signal integrity checks and circuit optimization. The proposed algorithm is verified with standard cells, simple digital circuits and ISCAS benchmark circuits in a 45nm technology. The results demonstrate the high accuracy and speed of our algorithm.

C Knoth, H Jedda and U Schlichtmann

This paper presents a new current source model (CSM) that allows to model noise on supply nets originating from CMOS logic cells. It also captures the influence of dynamic supply voltage changes on power consumption and cell delay. The CSM models n/pMOS blocks separately to reduce the complexity of model components. Compared with other CSMs, only two-dimensional tables are needed. This results in low characterization times and high simulation speed. Moreover, no re-characterization is needed for different supply voltages. The model is tested in a SPICE simulator. A reduction in transient simulation time by up to 53X was observed in the results, while the error in delay and current consumption was typically less than 3 percent.

[196] Clock Skew Scheduling for Timing Speculation [p. 929]
R Ye, F Yuan, H Zhou and Q Xu

By assigning intentional clock arrival times to the sequential elements in a circuit, clock skew scheduling (CSS) techniques can be utilized to improve IC performance. Existing CSS solutions work in a conservative manner that guarantees "always correct" computation, and hence their effectiveness is greatly challenged by the ever-increasing process variation effects. By allowing infrequent timing errors and recovering from
them with minor performance impact, timing speculation techniques such as Razor have gained wide interests from both academia and industry. In this work, we formulate the clock skew scheduling problem for circuits equipped with timing speculation capability and propose a novel CSS algorithm based on gradient-descent method. Experimental results on various benchmark circuits demonstrate the effectiveness of our proposed methodology.

8.6: System-Level Optimization of Embedded Real-Time Systems
Moderators: J Teich, Erlangen-Nuremberg U, DE; J-J Chen, Karlsruhe Institute of Technology, DE

Robust and Flexible Mapping for Real-time Distributed Applications during the Early Design Phases [p. 935]
J Gan, P Pop, F Gruian and J Madsen

We are interested in mapping hard real-time applications on distributed heterogeneous architectures. An application is modeled as a set of tasks, and we consider a fixed-priority preemptive scheduling policy. We target the early design phases, when decisions have a high impact on the subsequent implementation choices. However, due to a lack of information, the early design phases are characterized by uncertainties, e.g., in the worst-case execution times (wcets), or in the functionality requirements. We model uncertainties in the wcets using the "percentile method". The uncertainties in the functionality requirements are captured using "future scenarios", which are task sets that model functionality likely to be added in the future. In this context, we derive a mapping of tasks in the application, such that the resulted implementation is both robust and flexible. Robust means that the application has a high chance of being schedulable, considering the wcet uncertainties, whereas a flexible mapping has a high chance to successfully accommodate the future scenarios. We propose a Genetic Algorithm-based approach to solve this optimization problem. Extensive experiments show the importance of taking into account the uncertainties during the early design phases.

M A Bamakhrama, J T Zhai, H Nikolov and T Stefanov

The increasing complexity of modern embedded streaming applications imposes new challenges on system designers nowadays. For instance, the applications evolved to the point that in many cases hard-real-time execution on multiprocessor platforms is needed in order to meet the applications' timing requirements. Moreover, in some cases, there is a need to run a set of such applications simultaneously on the same platform with support for accepting new incoming applications at runtime. Dealing with all these new challenges increases significantly the complexity of system design. However, the design time must remain acceptable. This requires the development of novel systematic and automated design methodologies driven by the aforementioned challenges. In this paper, we propose such a novel methodology for automated design of an embedded multiprocessor system, which can run multiple hard-real-time streaming applications simultaneously. Our methodology does not need the complex and time-consuming design space exploration phase, present in most of the current state-of-the-art multiprocessor design frameworks. In contrast, our methodology applies very fast yet accurate schedulability analysis to determine the minimum number of processors, needed to schedule the applications, and the mapping of applications' tasks to processors. Furthermore, our methodology enables the use of hard-real-time multiprocessor scheduling theory to schedule the applications in a way that temporal isolation and a given throughput of each application are guaranteed. We evaluate an implementation of our methodology using a set of real-life streaming applications and demonstrate that it can greatly reduce the design time and effort while generating high quality hard-real-time systems.

K Jiang, P Eles and Z Peng

In this paper we consider distributed real-time embedded systems in which confidentiality of the internal communication is critical. We present an approach to efficiently implement cryptographic algorithms by using hardware/software co-design techniques. The objective is to find the minimal hardware overhead and corresponding process mapping for encryption and decryption tasks of the system, so that the confidentiality requirements for the messages transmitted over the internal communication bus are fulfilled, and time constraints are satisfied. Towards this, we formulate the optimization problems using Constraint Logic Programming (CLP), which returns optimal solutions. However, CLP executions are computationally expensive and, hence, efficient heuristics are proposed as an alternative. Extensive experiments demonstrate the efficiency of the proposed heuristic approaches.

8.7: On-Line Test for Secure Systems
Moderators: X Vera, Intel Labs Barcelona, ES; J Abella, Barcelona Supercomputing Center, ES

Logic Encryption: A Fault Analysis Perspective [p. 953]
J Rajendran, Y Pino, O Sinanoglu and R Kami
The globalization of Integrated Circuit (IC) design flow is making it easy for rogue elements in the supply chain to pirate ICs, overbuild ICs, and insert hardware trojans; the IC industry is losing approximately $4 billion annually [1], [2]. One way to protect the ICs from these attacks is to encrypt the design by inserting additional gates such that correct outputs are produced only when specific inputs are applied to these gates. The state-of-the-art logic encryption technique inserts gates randomly into the design [3] and does not necessarily ensure that wrong keys corrupt the outputs. We relate logic encryption to fault propagation analysis in IC testing and develop a fault analysis based logic encryption technique. This technique achieves 50% Hamming distance between the correct and wrong outputs (ideal case), when a wrong key is applied. Furthermore, this 50% Hamming distance target is achieved by using a smaller number of additional gates when compared to random logic encryption.

Low-Cost Implementations of On-the-Fly Tests for Random Number Generators [p. 959]
F Veljkovic, V Rozic and I Verbauwhede

Random number generators (RNG) are important components in various cryptographic systems. Embedded security systems often require a high-quality digital source of randomness. Still, randomness of an RNG can vary due to aging effects, temperature or process conditions or intentional active attacks. This paper presents efficient, compact and reliable hardware implementations of 8 tests from the NIST test suite for statistical evaluation of randomness. These tests can be used for on-the-fly quality monitoring of on-chip random number generators as well as for fast hardware evaluation of RNG designs.

Post-Deployment Trust Evaluation in Wireless Cryptographic ICs [p. 965]
Y Jin, D Maliuk and Y Makris

The use of side-channel parametric measurements along with statistical analysis methods for detecting hardware Trojans in fabricated integrated circuits has been studied extensively in recent years, initially for digital designs but recently also for their analog/RF counterparts. Such post-fabrication trust evaluation methods, however, are unable to detect dormant hardware Trojans which are activated after a circuit is deployed in its field of operation. For the latter, an on-chip trust evaluation method is required. To this end, we present a general architecture for post-deployment trust evaluation based on on-chip classifiers. Specifically, we discuss the design of an on-chip analog neural network which can be trained to distinguish trusted from untrusted circuit functionality based on simple measurements obtained via on-chip measurement acquisition sensors. The proposed method is demonstrated using a Trojan-free and two Trojan infested variants of a wireless cryptographic IC design, as well as a fabricated programmable neural network experimentation chip. As corroborated by the obtained experimental results, two current measurements suffice for the on-chip classifier to effectively assess trustworthiness and, thereby, detect hardware Trojans that are activated after chip deployment.

Batteries and Battery Management Systems for Electric Vehicles [p. 971]
M Brandl, H Gall, M Wenger, V Lorentz, M Giegerich, F Baronti, G Fantechi, L Fanucci, R Roncella, R Saletti, S Saponara, A Thaler, M Cifrain and W Prochazka

The battery is a fundamental component of electric vehicles, which represent a step forward towards sustainable mobility. Lithium chemistry is now acknowledged as the technology of choice for energy storage in electric vehicles. However, several research points are still open. They include the best choice of the cell materials and the development of electronic circuits and algorithms for a more effective battery utilization. This paper initially reviews the most interesting modeling approaches for predicting the battery performance and discusses the demanding requirements and standards that apply to ICs and systems for battery management. Then, a general and flexible architecture for battery management implementation and the main techniques for state-of-charge estimation and charge balancing are reported. Finally, we describe the design and implementation of an innovative BMS, which incorporates an almost fully-integrated active charge equalizer.

Keywords: Li-ion batteries, Cell Modeling, Battery Management System, Charge Equalization, State-of-Charge Estimation, Electric Vehicles.

Power Management of Multi-Core Chips: Challenges and Pitfalls [p. 977]

Modern processor systems are equipped with on-chip or on-board power controllers. In this paper, we examine the challenges and pitfalls in
architecting such dynamic power management control systems. A key question that we pose is: How to ensure that such managed systems are "energy-secure" and how to pursue pre-silicon modeling to ensure such security? In other words, we address the robustness and security issues of such systems. We discuss new advances in energy-secure power management, starting with an assessment of potential vulnerabilities in systems that do not address such issues up front.

P2012 is an area- and power-efficient many-core computing fabric based on multiple globally asynchronous, locally synchronous (GALS) clusters supporting aggressive fine-grained power, reliability and variability management. Clusters feature up to 16 processors and one control processor with independent instruction streams sharing a multi-banked L1 data memory, a multi-channel DMA engine, and specialized hardware for synchronization and scheduling. P2012 achieves extreme area and energy efficiency by supporting domain-specific acceleration at the processor and cluster level through the addition of dedicated HW IPs. P2012 can run standard OpenCL and OpenMP parallel codes as well as proprietary Native Programming Model (NPM) SW components that provide the highest level of control on application-to-resource mapping. In Q3 2011 the P2012 SW Development Kit (SDK) has been made available to a community of R&D users; it includes full OpenCL and NPM development environments. The first P2012 SoC prototype in 28nm CMOS will sample in Q4 2012, featuring four clusters and delivering 80GOPS (with single precision floating point support) in 18mm2 with 2W power consumption.

Personal health monitoring systems can offer a cost-effective solution for human healthcare. To extend the lifetime of health monitoring systems, we propose a near-threshold ultra-low-power multi-core architecture featuring low-power cores, yet capable of executing biomedical applications, with multiple instruction and data memories, tightly coupled through flexible crossbar interconnects. This architecture also includes broadcasting mechanisms for the data and instruction memories to optimize system energy consumption by tailoring memory sharing to the target application. Moreover, the architecture enables power gating of the unused memory banks to lower leakage power. Our experimental results show that compared to the state-of-the-art, the proposed architecture achieves 39.5% power savings at high workload requirements (637 MOps/s), and 38.8% savings at low workload requirements (5 kOps/s), whereby leakage power consumption dominates.

Future computing clusters will prevalently run parallel workloads to take advantage of the increasing number of cores on chips. In tandem, there is a growing need to reduce energy consumption of computing. One promising method for improving energy efficiency is co-scheduling applications on compute nodes. Efficient consolidation for parallel workloads is a challenging task as a number of factors, such as scalability, inter-thread communication patterns, or memory access frequency of the applications affect the energy/performance tradeoffs. This paper evaluates the impact of co-scheduling parallel workloads on the energy consumed per useful work done on real-life servers. Based on this analysis, we propose a novel multi-level technique that selects the best policy to co-schedule multiple workloads on a multi-core processor. Our measurements demonstrate that the proposed multi-level co-scheduling method improves the overall energy per work savings of the multi-core system up to 22% compared to state-of-the-art techniques.

9.3: Architecture and Building Blocks for Secure Systems

Moderators: L Fesquet, TIMA Laboratory, FR; L Torres, LIRMM, FR

Ensuring electronic components are free from Hardware Trojans is a very difficult task. Research suggests that even the best pre- and post-deployment detection mechanisms will not discover all malicious inclusions, nor prevent them from being activated. For economic reasons electronic components are used regardless of the possible presence of such Trojans. We developed the SAFER PATH architecture, which uses instruction and data fragmentation, program replication, and voting to create a computational system that is able to operate safely in the presence of active Hardware Trojans. We protect the integrity of the computation, the confidentiality of data being processed and ensure system availability. By combining a small Trusted Computing Base with Commercial-Off-The-Shelf processing elements, we are able to protect computation from the effects of arbitrary Hardware Trojans.
9.4: Advances in High-Level Synthesis

Moderators: G Coutinho, ICL, UK; P Coussy, Bretagne-Sud U, FR

ASIC Implementations of Five SHA-3 Finalists [p. 1006]

X Guo, M Srivastav, S Huang, D Ganta, M B Henry, L Nazhandali and P Schaumont

Throughout the NIST SHA-3 competition, in relative order of importance, NIST considered the security, cost, and algorithm and implementation characteristics of a candidate [1]. Within the limited one-year security evaluation period for the five SHA-3 finalists, the cost and performance evaluation may put more weight in the selection of winner. This work contributes to the SHA-3 hardware evaluation by providing timely cost and performance results on the first SHA-3 ASIC in 0.13μm IBM process using standard cell CMOS technology with measurements of all the five finalists using the latest Round 3 tweaks. This article describes the SHA-3 ASIC design from VLSI architecture implementation to the silicon realization.

Side Channel Analysis of the SHA-3 Finalists [p. 1012]

M Zohner, M Kasper, M Stoettinger and S A Huss

At the cutting edge of today's security research and development, the SHA-3 competition evaluates a new secure hashing standard in succession to SHA-2. The five remaining candidates of the SHA-3 competition are BLAKE, Grøstl, JH, Keccak, and Skein. While the main focus was on the algorithmic security of the candidates, a side channel analysis has only been performed for BLAKE and Grøstl [1]. In order to equally evaluate all candidates, we identify side channel attacks on JH-MAC, Keccak-MAC, and Skein-MAC and demonstrate the applicability of the attacks by attacking their respective reference implementation. Additionally, we revisit the side channel analysis of Grøstl and introduce a profiling based side channel attack, which emphasizes the importance of side channel resistant hash functions by recovering the input to the hash function using only the measured power consumption.

Index Terms - SHA-3 Finalists; Side-Channel Analysis; DPA

Combining Module Selection and Replication for Throughput-Driven Streaming Programs [p. 1018]

J Cong, M Huang, B Liu, P Zhang and Y Zou

Streaming processing is widely adopted in many data-intensive applications in various domains. FPGAs are commonly used to realize these applications since they can exploit inherent data parallelism and pipelining in the applications to achieve a better performance. In this paper we investigate the design space exploration problem (DSE) when mapping streaming applications onto FPGAs. Previous works narrowly focus on using techniques like replication or module selection to meet the throughput target. We propose to combine these two techniques together to guide the design space exploration. A formal formulation and solution to this combined problem is presented in this paper. Our objective is to optimize the total area cost subject to the throughput constraint. In particular, we are able to handle the feedback loops in the streaming programs, which, to the best of our knowledge, has never been discussed in previous work. Our methodology is evaluated with high-level synthesis tools, and we demonstrate our workflow on a set of benchmarks that vary from module kernel design such as FFT to large designs such as an MPEG-4 decoder.

Exploiting Area/Delay Tradeoffs in High-Level Synthesis [p. 1024]

A Kondratyev, L Lavagno, M Meyer and Y Watanabe

This paper proposes an enhanced scheduling approach for high-level synthesis, which relies on a multi-cycle behavioral timing analysis step that is performed before and during scheduling. The goal of this analysis is to accurately evaluate the criticality of operations and determine the most suitable candidate resources to implement them. The efficiency of the approach is confirmed by testing it on industrial examples, where it achieves, on average, 9% area savings after logic synthesis.

Predicting Best Design Trade-offs: A Case Study in Processor Customization [p. 1030]

M Zuluaga, E Bonilla and N Topham

Given the high level description of a task, many different hardware modules may be generated while meeting its behavioral requirements. The characteristics of the generated hardware can be tailored to favor energy efficiency, performance, accuracy or die area. The inherent trade-offs between such metrics need to be explored in order to choose a solution that meets design and cost expectations. We address the generic problem of automatically deriving a hardware implementation from a high-level task description. In this paper we present novel technique that exploits previously explored implementation design spaces in order to find optimal trade-offs for new high-level descriptions. This technique is generalizable to a range high-level synthesis problems in which trade-offs can be exposed by changing the parameters of the hardware generation tool. Our strategy, based upon machine learning techniques, models the impact of the parameterization of the tool on the target
9.5: Supply Voltage and Circuitry Based Power Reductions

Moderators: M Lopez-Vallejo, UP Madrid, ES; W Nebel, Oldenburg U and OFFIS, DE

Objectives, given the characteristics of the input. Thus, a predictor is able suggest a subset of parameters that are likely to lead to optimal hardware implementations. The proposed method is evaluated on a resource sharing problem which is typical in high level synthesis, where the trade-offs between area and performance need to be explored. In this case study, we show that the technique can reduce by two orders of magnitude the number of design points that need to be explored in order to find the Pareto optimal solutions.

R Wille, R Drechsler, C Osewold and A Garcia-Ortiz

The application of coding strategies is an established methodology to improve the characteristics of on-chip interconnect architectures. Therefore, design methods are required which realize the corresponding encoders and decoders with as small as possible overhead in terms of power and delay. In the past, conventional design methods have been applied for this purpose. This work proposes an entirely new direction which exploits design methods for reversible circuits. Here, much progress has been made in the last years. The resulting reversible circuits represent one-to-one mappings which can inherently work logical descriptions for the desired encoders and decoders. Both, an exact and a heuristic synthesis approach, are introduced which rely on reversible design principles but also incorporate objectives from on-chip interconnect architectures. Experiments show that significant improvements with respect to power consumption, area, and delay can be achieved using the proposed direction.

[215] Ultra Low Power Litho Friendly Local Assist Circuitry for Variability Resilient 8T SRAM [p. 1042]
V Sharma, S Cosemans, M Ashouei, J Huisken, F Catthoor and W Dehaene

This paper presents litho friendly circuit techniques for variability resilient low power 8T SRAM. The new local assist circuitry achieves a state-of-the-art low energy and variability resilient WRITE operation and improves the degraded access speed of SRAM cells at low voltages. Differential VSS bias increases the variability resilience. The physical regularity in the layout of local assist circuitry enables litho optimization thereby reducing the area overhead associated with existing local assist techniques. Statistical simulations in 40nm LP CMOS technology reveals 10x reduction in WRITE energy consumption, 103x reduction in write failures, 6.5x improvement in read access time and 31% reduction in the area overhead.

Keywords- SRAM 8T cell, variation, Write Margin, local write receiver, litho optimized.

[216] Sliding-Mode Control to Compensate PVT Variations in Dual Core Systems [p. 1048]
H R Pourshaghaghi, H Fatemi and J Pineda de Gyvez

In this paper, we present a novel robust sliding-mode controller for stabilizing supply voltage and clock frequency of dual core processors determined by dynamic voltage and frequency scaling (DVFS) methods in the presence of systematic and random variations. We show that maximum rejection for process, voltage and temperature (PVT) variations can be achieved by using the proposed sliding-mode controller. The stabilization of the presented controller is confirmed by the Lyapunov method. Experimental results demonstrate maximum 20% robustness against 20% parameter variations for a hardware of two core processors executing a JPEG decoding application.

Keywords- Sliding-Mode Feedback Control; PVT Variations; Systematic and Random Variations; Dual Core System;

K Jeong, A B Kahng, S Kang, T S Rosing and R Strong

In mobile systems, the problems of short battery life and increased temperature are exacerbated by wasted leakage power. Leakage power waste can be reduced by power-gating a core while it is stalled waiting for a resource. In this work, we propose and model memory access power gating (MAPG), a low-overhead technique to enable power gating of an active core when it stalls during a long memory access. We describe a programmable two-stage power gating switch design that can vary a core’s wake-up delay while maintaining voltage noise limits and leakage power savings. We also model the processor power distribution network and the effect of memory access power gating on neighboring cores. Last, we apply our power gating technique to actual benchmarks, and examine energy savings and overheads from power gating stalled cores during long memory accesses. Our analyses show the potential for over 38% energy savings given “perfect” power gating on memory accesses; we achieve energy savings exceeding 20% for a practical, counter-based implementation.

Q Xie, X Lin, Y Wang, M Pedram, D Shin and N Chang
This paper is the first to present an efficient charge management algorithm focusing on extending the cycle life battery elements in hybrid electrical energy storage (HEES) systems while simultaneously improving the overall cycle efficiency. In particular, it proposes to apply a crossover filter the power source and load profiles. The goal of this filtering technique is to allow the battery banks to stably (i.e., with low variation) receive energy from the power source and/or provide energy to the load device, while leaving the spiky (i.e., with high variation) power supply or demand to be dealt with by the supercapacitor banks. To maximize the HEES system cycle efficiency, a mathematical problem is formulated and solved to determine the optimal charging/discharging current profiles and charge transfer interconnect voltage, taking into account the power loss of the EES elements and power converters. To minimize the state of health (SoH) degradation of the battery array in the HEES system, we make use of two facts: the SoH battery is better maintained if (i) the SoC swing is smaller, and (ii) the same SoC swing occurs at lower average SoC. Now then using the supercapacitor bank to deal with the high-frequency component of the power supply or demand, we can reduce the SoC swing for the battery array and lower the SoC of the array. A secondary helpful effect is that, for fixed and given amount energy delivered to the load device, an improvement in the overall charge cycle efficiency of the HEES system translates into a further reduction in both the average SoC and the SoC swing the battery array. The proposed charge management algorithm for a Li-ion battery - supercapacitor bank HEES system simulated and compared to a homogeneous EES system comprised of Li-ion batteries only. Experimental results show significant performance enhancements for the HEES system, an increase of up to 21.9% and 4.82x in terms of the cycle efficiency and cycle life, respectively.

Keywords: hybrid electrical energy storage system, charge management, state of health.

9.6: Creation and Processing of System-level Models
Moderators: E Villar, Cantabria U, ES; J Haase, TU Wien, AT

V Todorov, D Mueller-Gritschneder, H Reinig and U Schlichtmann

Transaction level (TL) models are key to early design exploration, performance estimation and virtual prototyping. Their speed and accuracy enable early and rapid System-on-Chip (SoC) design evaluation and software development. Most devices have only register transfer level (RTL) models that are too complex for SoC simulation. Abstracting these models to TL ones, however, is a challenging task, especially when the RTL description is too obscure or not accessible. This work presents a methodology for automatically creating a TL model of an RTL memory controller component. The device is treated as a black box and a multitude of simulations is used to obtain results, showing its timing behavior. The results are classified into conditional probability distributions, which are reused within a TL model to approximate the RTL timing behavior. The presented method is very fast and highly accurate. The resulting TL model executes approximately 1200 times faster, with maximum measured average timing offset error of 7.66%.

E Ebeid, F Fummi, D Quaglia and F Stefanni

Network design in distributed embedded applications is a novel challenging task which requires 1) the extraction of communication requirements from application specification and 2) the choice of channels and protocols connecting physical nodes. These issues are faced in the paper by adopting UML/MARTE as specification front-end and repository of refined versions of the model obtained by both simulation and analytical exploration of the design space. The emphasis is on using standard UML/MARTE elements for the description of networked embedded systems to allow re-use, tool interoperability and documentation generation. The approach is explained on a case study related to building automation.

[221] Debugging of Inconsistent UML/OCL Models [p. 1078]
R Wille, M Soeken and R Drechsler

While being a de-facto standard for the modeling of software systems, the Unified Modeling Language (UML) is also increasingly used in the domain of hardware design and hardware/software co-design. To ensure the correctness of the specified systems, approaches have been presented which automatically verify whether a UML model is consistent, i.e. free of conflicts. However, if the model is inconsistent, these approaches do not provide further information to assist the designer in finding the error. In this work, we present an automatic debugging approach which determines contradiction candidates, i.e. a small subset of the original model explaining the conflict. These contradiction candidates aid the designer in finding the error faster and therefore accelerate the whole design process. The approach employs different satisfiability solvers as well as different debugging strategies. Experimental results demonstrate that, even for large UML models with up to 2500 classes and constraints, the approach determines a very small number of contradiction candidates to be inspected.

9.7: Test and Monitoring of RF and Mixed-Signal ICs
Moderators: S Sattler, Erlangen-Nuremberg U, DE; H Stratigopoulos, IMAG/CNRS, FR
An Analytical Technique for Characterization of Transceiver IQ Imbalances in the Loop-Back Mode [p. 1084]  
A Nassery and S Ozev

Loop-back is a desirable test set-up for RF transceivers for both on-chip characterization and production testing. Measurement of IQ imbalances (phase mismatch, gain mismatch, DC offset, and time skews) in the loop-back mode is challenging due to the coupling between the receiver (RX) and transmitter (TX) parameters. We present an analytical method for the measurement of the imbalances in the loop-back mode. We excite the system with carefully designed test signals at the baseband TX input and analyze the corresponding RX baseband output. The derived and used mathematical equations based on these test inputs enable us to unambiguously compute IQ imbalances. Experiments conducted both in simulations and on a hardware platform confirm that the proposed technique can accurately compute the IQ imbalances.

Testing RF Circuits with True Non-Intrusive Built-In Sensors [p. 1090]  
L Abdallah, H-G Stratigopoulos, S Mir and J Altet

We present a set of sensors that enable a built-in test in RF circuits. The key characteristic of these sensors is that they are non-intrusive, that is, they are not electrically connected to the RF circuit, and, thereby, they do not degrade its performances. In particular, the presence of spot defects is detected by a temperature sensor, whereas the performances of the RF circuit in the presence of process variations are implicitly predicted by process sensors, namely dummy circuits and process control monitors. We discuss the principle of operation of these sensors, their design, as well as the test strategy that we have implemented. The idea is demonstrated on an RF low noise amplifier using post-layout simulations.

Monitoring Active Filters under Automotive Aging Scenarios with Embedded Instrument [p. 1096]  
J Wan and H G Kerkhoff

in automotive mixed-signal SoCs, the analogue/mixed-signal front-ends are of particular interest with regard to dependability. Because of the many electrical disturbances at the front-end, often (active) filters are being used. Due to the harsh environments, in some cases, degradation of these filters may be encountered during lifetime and hence false sensor information could be provided with potential fatal results. This paper investigates the influence of aging in three different types of active filters in an automotive environment, and presents an embedded instrument, which monitors this aging behaviour. The monitor can be used for flagging problems in the car console or initiate automatic correction. Keywords-component; active filters; testing; aging; monitoring; NBTI; embedded instruments

Analysis of Instruction-level Vulnerability to Dynamic Voltage and Temperature Variations [p. 1102]  
A Rahimi, L Benini and R K Gupta

Variation in performance and power across manufactured parts and their operating conditions is an accepted reality in aggressive CMOS processes. This paper considers challenges and opportunities in identifying this variation and methods to combat it for improved computing systems. We introduce the notion of instruction-level vulnerability (ILV) to expose variation and its effects to the software stack for use in architectural/compiler optimizations. To compute ILV, we quantify the effect of voltage and temperature variations on the performance and power of a 32-bit, RISC, in-order processor in 65nm TSMC technology at the level of individual instructions. Results show 3.4ns (68FO4) delay variation and 26.7x power variation among instructions, and across extreme corners. Our analysis shows that ILV is not uniform across the instruction set. In fact, ILV data partitions instructions into three equivalence classes. Based on this classification, we show how a low-overhead robustness enhancement techniques can be used to enhance performance by a factor of 1.1x-5.5x.

CrashTesting SWAT: Accurate, Gate-Level Evaluation of Symptom-Based Resiliency Solutions [p. 1106]  
A Pellegrini, R Smolinski, L Chen, X Fu, S K S Hari, J Jiang, S V Adve, T Austin and V Bertacco

Current technology scaling is leading to increasingly fragile components, making hardware reliability a primary design consideration. Recently researchers have proposed low-cost reliability solutions that detect hardware faults through software-level symptom monitoring. SWAT (SoftWare Anomaly Treatment), one such solution, demonstrated with microarchitecture-level simulations that symptom-based solutions can provide high fault coverage and a low Silent Data Corruption (SDC) rate. However, more accurate evaluations are needed to validate such solutions for hardware faults in real-world processor designs. In this paper, we evaluate SWAT's symptom-based detectors on gate-level faults using an FPGA-based, full-system prototype. With this platform, we performed a gate-level accurate fault injection campaign of 51,630 fault injections in the OpenSPARC T1 core logic across five SPECint 2000 benchmarks. With an overall SDC rate of 0.79%, our results are
comparable to previous microarchitecture-level evaluations of SWAT, demonstrating the effectiveness of symptom-based software detectors for permanent faults in real-world designs.

A Hybrid HW-SW Approach for Intermittent Error Mitigation in Streaming-Based Embedded Systems [p. 1110]
M M Sabry, D Atienza and F Catthoor

Recent advances in process technology augment the systems-on-chip (SoCs) functionality per unit area with the substantial decrease of device features. However, features abatement triggers new reliability issues such as the single-event multi-bit upset (SMU) failure rates augmentation. To mitigate these failure rates, we propose a novel error mitigation mechanism that relies on a hybrid HW-SW technique. In our proposal, we enforce SoC SRAMs by implementing a fault-tolerant memory buffer with minimal capacity to ensure error-free operation. We utilize this buffer to temporarily store a portion of the stored data, named a data chunk, that is used to restore another data chunk in a fully demand-driven way, in case the latter is faulty. We formulate the buffer and data chunk size selection as an optimization problem that targets energy overhead minimization, given that timing and area overheads are restricted with hard constraints decided beforehand by the system designers. We show that our proposed mitigation scheme achieves full error mitigation in a real SoC platform with an average of 10.1% energy overhead with respect to a base-line system operation, while guaranteeing all the design-time constraints.

Probabilistic Response Time Bound for CAN Messages with Arbitrary Deadlines [p. 1114]
P Axer, M Sebastian and R Ernst

The controller area network (CAN) is widely used in industrial and the automotive domain and in this context often for hard real-time applications. Formal methods guide the designer to give worst-case guarantees on timing. However, due to bit errors on the communication channel response times can be delayed due to retransmissions. Some methods exist to cover these effects, but are limited e.g. (support only periodic real-time traffic). In this paper we generalize existing methods to support arbitrary deadlines, and derive a probabilistic response time bound which is especially useful with the emergence of the new automotive safety standard ISO 26262.

Exploring Pausible Clocking Based GALS Design for 40-nm System Integration [p. 1118]
X Fan, M Kristic, E Grass, B Sanders and C Heer

Globally asynchronous locally synchronous (GALS) design has attracted intensive research attention during the last decade. Among the existing GALS design solutions, the pausible clocking scheme presents an elegant solution to address the cross-clock synchronization issues with low hardware overhead. This work explored the applications of pausible clocking scheme for area/power efficient GALS design. To alleviate the challenge of timing convergence at the system level, area and power balanced system partitioning was applied for GALS design. An optimized GALS design flow based on the pausible clocking scheme was further proposed. As a practical example, a synchronous/GALS OFDM baseband transmitter chip, named Moonrake, was then designed and fabricated using the 40-nm CMOS process. It is shown that, compared to the synchronous baseline design, 5% reduction in area and 6% saving in power can be achieved in the GALS counterpart.

Keywords - SoC, GALS, pausible clocking, OFDM

Static Analysis of Asynchronous Clock Domain Crossings [p. 1122]
S Chaturvedi

Clock domain crossing (CDC) signals pose unique and challenging issues in complex designs with multiple asynchronous clocks running at frequencies as high as multiple giga hertz. Designers can no longer rely on ad hoc approaches to CDC analysis. This paper describes a methodical approach for static analysis of structural issues in asynchronous CDCs. The illustrated approach can be integrated easily in standard static timing analysis (STA) flows of any design house. The methodology was successfully deployed on a 32 nm accelerated processing unit (APU) design, and a case study of the same is included in this paper.

Keywords - STA, CDC, clock domain crossing

A Scalable GPU-based Approach to Accelerate the Multiple-Choice Knapsack Problem [p. 1126]
B Suri, U D Bordoloi and P Eles

Variants of the 0-1 knapsack problem manifest themselves at the core of several system-level optimization problems. The running times of such system-level optimization techniques are adversely affected because the knapsack problem is NP-hard. In this paper, we propose a new GPU-based approach to accelerate the multiple-choice knapsack problem, which is a general version of the 0-1 knapsack problem. Apart from exploiting the parallelism offered by the GPUs, we also employ a variety of GPU-specific optimizations to further accelerate the running times of the knapsack problem. Moreover, our technique is scalable in the sense that even when running large instances of the multiple-choice knapsack problems, we can efficiently utilize the GPU compute resources and memory bandwidth to achieve significant speedups.
Enhancing Non-Linear Kernels by an Optimized Memory Hierarchy in a High Level Synthesis Flow [p. 1130]

S Mancini and F Rousseau

Modern High Level Synthesis (HLS) tools are now efficient at generating RTL models from algorithmic descriptions of the target hardware accelerators but they still do not manage memory hierarchies. Memory hierarchies are efficiently optimized by performing code transformations prior to HLS in frameworks which exploit the linearity of the mapping functions between loop indexes and memory references (called linear kernels). Unfortunately, non-linear kernels are algorithms which do not benefit of such classical frameworks, because of the disparity of the non-linear functions to compute their memory references. In this paper we propose a method to design non-linear kernels in a HLS flow, which can be seen as a code pre-processing. The method starts from an algorithmic description and generates an enhanced algorithmic description containing both the non-linear kernel and an optimized memory hierarchy. The transformation and the associated optimization process provides a significant gain when compared to a standard optimization. Experiments on benchmarks show an average reduction of 28% of the external memory traffic and about 32 times of the embedded memory size.

Workload-Aware Voltage Regulator Optimization for Power Efficient Multi-Core Processors [p. 1134]

A A Sinkar, H Wang and N S Kim

Modern multi-core processors use power management techniques such as dynamic voltage and frequency scaling (DVFS) and clock gating (CG) which cause the processor to operate in various performance and power states depending on runtime workload characteristics. A voltage regulator (VR), which is designed to provide power to the processor at its highest performance level, can significantly degrade in efficiency when the processor operates in the deep power saving states. In this paper, we propose VR optimization techniques to improve the energy efficiency of the processor + VR system by using the workload dependent P- and C-state residency of real processors. Our experimental results for static VR optimization show up to 19%, 20%, and 4% reduction in energy consumption for workstation, mobile and server multi-core processors. We also investigate the effect of dynamically changing VR parameters on the energy efficiency compared to the static optimization.

Keywords-DVFS; switching voltage regulator; P-state; C-state;

An Energy Efficient DRAM Subsystem for 3D Integrated SoCs [p. 1138]

C Weis, I Loi, L Benini and N Wehn

Energy efficiency is the key driver for the design optimization of System-on-Chips for mobile terminals (smartphones and tablets). 3D integration of heterogeneous dies based on TSV (through silicon via) technology enables stacking of multiple memory or logic layers and has the advantage of higher bandwidth at lower energy consumption for the memory interface. In this work we propose a highly energy efficient DRAM subsystem for next-generation 3D integrated SoCs, which will consist of a SDR/DDR 3D-DRAM controller and an attached 3D-DRAM cube with a fine-grained access and a very flexible (WIDE-IO) interface. We implemented a synthesizable model of the SDR/DDR 3-DRAM channel controller and a functional model of the 3D-stacked DRAM which embeds an accurate power estimation engine. We investigated different DRAM families (WIDE-IO DDR/SDR, LPDDR and LPDDR2) and densities that range from 256Mb to 4Gb per channel. The implementation results of the proposed 3D-DRAM subsystem show that energy optimized accesses to the 3D-DRAM enable an overall average of 37% power savings as compared to standard accesses. To the best of our knowledge this is the first design of a 3D-DRAM channel controller and 3D-DRAM model featuring co-optimization of memory and controller architecture.

Eliminating Invariants in UML/OCL Models [p. 1142]

M Soeken, R Wille and R Drechsler

In model-based design, it is common and helpful to use invariants in order to highlight restrictions or to formulate characteristics of a design. In contrast to pre- and post-conditions, they represent global constraints. That is, they are harder to explicitly consider and, thus, become disadvantageous when the design process approaches the implementation phase. As a consequence, they should be removed from a design when it comes to an implementation. However, so far only naive tool support aiding the designer in this task is available. In this paper, we present an approach which addresses this problem. A methodology is proposed which iteratively removes invariants from a model and, afterwards, presents the designer with invalid scenarios originally prevented by the just eliminated invariant. Using this, the designer can either manually modify the model or simply take the automatically generated suggestion. This enables to entirely eliminate all invariants without changing the semantics of the model. Case studies illustrate the applicability of the proposed approach.

On-Chip Source Synchronous Interface Timing Test Scheme with Calibration [p. 1146]

H Kim and J A Abraham

This paper presents an on-chip test circuit with a high resolution for testing source synchronous interface timing. Instead of a traditional strobe-
scanning method, an on-chip delay measurement technique which detects the timing mismatches between data and clock paths is developed. Using a programmable pulse generator, the timing mismatches are detected and converted to pulse widths. To obtain digital test results compatible with low-cost ATE, an Analog-to-Digital Converter (ADC) is used. We propose a novel calibration method for the input range for the ADC using a binary search algorithm. This enables test results to be measured with high resolution using only a 4-bit flash ADC (which keeps the area overhead low). The method achieves a resolution of 21.88 ps in 0.18μ technology. We also present simulation results of the interface timing characterization, including timing margins and timing pass/fail decisions.

Keywords - Source-Synchronous, Memory Interfaces, ATE, Delay Measurement, Flash ADC, Calibration

10.1: Special Day More-than-Moore: Technologies

Moderators: M Brillouët, CEA-Leti, FR


H Graeb

In its recent 2011 version, The Technology Roadmap for Semiconductors [1] updated a section on analog design technology challenges. In the paper at hand, these challenges and exemplary solution approaches will be sketched. In detail, structure and symmetry analysis, analog placement, design for aging, discrete sizing, sizing with in-loop layout, and performance space exploration will be touched.

Keywords - analog design, placement, layout, sizing, reliability, aging, yield, Pareto, optimization, synthesis

10.2: Pathways to Servers of the Future

Moderator: G Fettweis, TU Dresden, DE


D Morche, M Pelissier, G Masson and P Vincent

This work presents the potential offered by new UWB pulse radio transceiver designs. It shows that a judicious architecture selection can be used to exploit the benefit of impulse radio and to reach state of the art performances both in energy efficiency and ranging accuracy. The first presented architecture is dedicated to localization application whereas the second one is focusing on high speed and remote powered radio link for the application of ambient intelligent. After a brief description of the application area, the selected architecture are described and justified. Then, the chipset design is presented and the measurements results are summarized. Lastly, perspectives are drawn from the combination of those two developments.

Keywords - Impulse Radio, UWB Transceiver, Ranging, double quadrature receiver, RFID, Ambient Intelligence, Memory Tag, super-regenerative oscillator

10.3: Side-Channel Analysis and Protection of Secure Embedded Systems

Moderators: F Regazzoni, ALaRI, CH; P Schaumont, Virginia Tech, US

[240] Amplitude Demodulation-based EM Analysis of Different RSA Implementations [p. 1167]

G Perin, L Torres, P Benoit and P Maurine

This paper presents a fully numeric amplitude-demodulation based technique to enhance simple electromagnetic analyses. The technique, thanks to the removal of the clock harmonics and some noise sources, allows efficiently disclosing the leaking information. It has been applied to three different modular exponentiation algorithms, mapped onto the same multiplexed architecture. The latter is able to perform the exponentiation with successive modular multiplications using the Montgomery method. Experimental results demonstrate the efficiency of the applied demodulation based technique and also point out the remaining weaknesses of the considered architecture to retrieve secret keys.

Keywords - Public-Key Cryptography, RSA, Modular Exponentiation, Side-Channel Attacks, AM Demodulation.
Amongst the many existing countermeasures against Side Channel Attacks (SCA) on symmetrical cryptographic algorithms, masking is one of the most widespread, thanks to its relatively low overhead, its low performance loss and its robustness against first-order attacks. However, several articles have recently pinpointed the limitations of this countermeasure when matched with variance-based and other high-order analyses. In this article, we present a new form of Boolean masking for the Advanced Encryption Standard (AES) called “RSM”, which shows the same level in performances as the state-of-the-art, while being less area consuming, and secure against Variance-based Power Analysis (VPA) and second-order zero-offset CPA. Our theoretical security evaluation is then validated with simulations as well as real-life CPA and VPA on an AES 256 implemented on FPGA.

Keywords: Side-Channel Attacks (SCA), Variance-based Power Analysis (VPA), zero-offset DPA, Mutual Information Analysis (MIA), substitution boxes (S-Boxes), Advanced Encryption Standard (AES), Boolean masking, Rotating S-boxes Masking (RSM).

In the light of implementation attacks a better understanding of complex circuits of security sensitive applications is an important issue. Appropriate evaluation tools and metrics are required to understand the origin of implementation flaws within the design process. The selected leakage model has significant influence on the reliability of evaluation results concerning the side-channel resistance of a cryptomorphic implementation. In this contribution we introduce methods, which determine the accuracy of the leakage characterization and allow to quantify the signal-to-noise ratio. This allows a quantitative assessment of the side-channel resistance of an implementation without launching an attack. We validate the conclusions drawn from our new methods by real attacks and obtain similar results. Compared to the commonly used Hamming Distance model in our experiments enhanced leakage models increased the attack efficiency by up to 500%.

Keywords: signal-to-noise ratio, approximation error, constructive side-channel analysis, secure hardware design

Three-dimensional (3D) circuit integration is a promising technology to alleviate performance and power related issues raised by interconnects in nanometer CMOS. Physical planning of three-dimensional integrated circuits is substantially different from that of traditional planar integrated circuits, due to the presence of multiple layers of dies. To realize the full potential offered by three-dimensional integration, it is necessary to take physical information into consideration at higher-levels of the design abstraction for 3D ICs. This paper proposes an incremental system-level synthesis framework that tightly integrates behavioral synthesis of modules into the layer assignment and floorplanning stage of 3D IC design. Behavioral synthesis is implemented as a sub-routine to be called to adjust delay/power/variability/area of circuit modules during the physical planning process. Experimental results show that with the proposed synthesis during planning methodology, the overall timing yield is improved by 8%, and the chip peak temperature reduced by 6.6 °C, compared to the conventional planning-after-synthesis approach.

This paper introduces the first exact method for optimal resource sharing in a pipelined system in order to minimize area. Given as input a dependence graph and a throughput requirement, our approach searches through the space of legal resource allocations, performing both scheduling and optimal buffer insertion, in order to produce the minimum area implementation. Furthermore, we do not arbitrarily limit the number of concurrent threads or data tokens; instead, we explore the full space of legal token counts, effectively allowing the depth of pipelining to be determined by our algorithm, while concurrently minimizing area and meeting performance constraints. Our approach has been automated, and compared with an existing single-token scheduling approach. Experiments using a set of benchmarks indicate that our multi-token approach has significant advantages: (i) it can find schedules that deliver higher throughput than the single-token approach; and (ii) for the same throughput, the multi-token approach obtains solutions that consumed 33-61% less area.

Three-dimensional (3D) circuit integration is a promising technology to alleviate performance and power related issues raised by interconnects in nanometer CMOS. Physical planning of three-dimensional integrated circuits is substantially different from that of traditional planar integrated circuits, due to the presence of multiple layers of dies. To realize the full potential offered by three-dimensional integration, it is necessary to take physical information into consideration at higher-levels of the design abstraction for 3D ICs. This paper proposes an incremental system-level synthesis framework that tightly integrates behavioral synthesis of modules into the layer assignment and floorplanning stage of 3D IC design. Behavioral synthesis is implemented as a sub-routine to be called to adjust delay/power/variability/area of circuit modules during the physical planning process. Experimental results show that with the proposed synthesis during planning methodology, the overall timing yield is improved by 8%, and the chip peak temperature reduced by 6.6 °C, compared to the conventional planning-after-synthesis approach.

This paper introduces the first exact method for optimal resource sharing in a pipelined system in order to minimize area. Given as input a dependence graph and a throughput requirement, our approach searches through the space of legal resource allocations, performing both scheduling and optimal buffer insertion, in order to produce the minimum area implementation. Furthermore, we do not arbitrarily limit the number of concurrent threads or data tokens; instead, we explore the full space of legal token counts, effectively allowing the depth of pipelining to be determined by our algorithm, while concurrently minimizing area and meeting performance constraints. Our approach has been automated, and compared with an existing single-token scheduling approach. Experiments using a set of benchmarks indicate that our multi-token approach has significant advantages: (i) it can find schedules that deliver higher throughput than the single-token approach; and (ii) for the same throughput, the multi-token approach obtains solutions that consumed 33-61% less area.
The multiple constant multiplications (MCM) operation, which realizes the multiplication of a set of constants by a variable, has a significant impact on the complexity and performance of the digital finite impulse response (FIR) filters. Over the years, many high-level algorithms and design methods have been proposed for the efficient implementation of the MCM operation using only addition, subtraction, and shift operations. The main contribution of this paper is the introduction of a high-level synthesis algorithm that optimizes the area of the MCM operation and, consequently, of the FIR filter design, on field programmable gate arrays (FPGAs) by taking into account the implementation cost of each addition and subtraction operation in terms of the number of fundamental building blocks of FPGAs. It is observed from the experimental results that the solutions of the proposed algorithm yield less complex FIR filters on FPGAs with respect to those whose MCM part is implemented using prominent MCM algorithms and design methods.

10.5: Modeling of Complex Analogue and Digital Systems
Moderators: T Kazmierski, Southampton U, UK; N van der Meijs, TU Delft, NL

10.6: Cyber-Physical Systems
Moderators: P Eles, Linkoping U, SE; R Ernst, TU Braunschweig, DE
We present an automatic schedule synthesis framework for applications that are mapped onto distributed time-triggered automotive platforms where multiple Electronic Control Units (ECUs) are synchronized over a FlexRay bus. We classify applications into two categories: (i) safety-critical control applications with stability and performance constraints, and (ii) time-critical applications with only deadline constraints. Our proposed framework can handle such mixed constraints arising from timing, control stability, and performance requirements. In particular, we synthesize schedules that optimize control performance and respects the timing requirements of the real-time applications. An Integer Linear Programming (ILP) problem is formulated by modeling the ECU and bus schedules as a set of constraints for optimizing both linear or quadratic control performance functions.

A Masrur, D Goswami, S Chakraborty, J-J Chen, A Annaswamy and A Banerjee

Many cyber-physical systems consist of a collection of control loops implemented on multiple electronic control units (ECUs) communicating via buses such as FlexRay. Such buses support hybrid communication protocols consisting of a mix of time- and event-triggered slots. The time-triggered slots may be perfectly synchronized to the ECUs and hence result in zero communication delay, while the event-triggered slots are arbitrated using a priority-based policy and hence messages mapped onto them can suffer non-negligible delays. In this paper, we study a switching scheme where control messages are dynamically scheduled between the time-triggered and the event-triggered slots. This allows more efficient use of time-triggered slots which are often scarce and therefore should be used sparingly. Our focus is to perform a schedulability analysis for this setup, i.e., in the event of an external disturbance, can a message be switched from an event-triggered to a time-triggered slot within a specified deadline? We show that this analysis can check whether desired control performance objectives may be satisfied, with a limited number of time-triggered slots being used.

Y Luo, K Chakrabarty and T-Y Ho

Droplet-based “digital” microfluidics technology has now come of age and software-controlled biochips for healthcare applications are starting to emerge. However, today's digital microfluidic biochips suffer from the drawback that there is no feedback to the control software from the underlying hardware platform. Due to the lack of precision inherent in biochemical experiments, errors are likely during droplet manipulation, but error recovery based on the repetition of experiments leads to wastage of expensive reagents and hard-to-prepare samples. By exploiting recent advances in the integration of optical detectors (sensors) in a digital microfluidics biochip, we present a “physical-aware” system reconfiguration technique that uses sensor data at checkpoints to dynamically reconfigure the biochip. A re-synthesis technique is used to recompute electrode-actuation sequences, thereby deriving new schedules, module placement, and droplet routing pathways, with minimum impact on the time-to-response.

[253] Predictive Control of Networked Control Systems over Differentiated Services Lossy Networks [p. 1245]
R Muradore, D Quaglia and P Fiorini

Networked control systems are feedback systems where plant and controller are connected through lossy wired/wireless networks. To mitigate communication delays and packet losses different control solutions have been proposed. In this work the model predictive control (MPC) has been improved by introducing transmission options offering different probabilities of packet drops (high priority service and low priority service). This Differentiated Services architecture introduces Quality-of-Service (QoS) guarantees and can be used to jointly design the control command and the transmission strategy. A novel MPC-QoS controller is proposed and its design is obtained by solving a mixed integer quadratic problem.

10.7: On-Line Test and Fault Tolerance
Moderators: D Gizopoulos, Athens U, GR; M Nicolaidis, TIMA Laboratory, FR

[254] Input Vector Monitoring on Line Concurrent BIST Based on Multilevel Decoding Logic [p. 1251]
I Voyiatzis

Input Vector Monitoring Concurrent Built-In Self Test (BIST) schemes provide the capability to perform testing while the Circuit Under Test (CUT) operates normally, by exploiting vectors that appear at the inputs of the CUT during its normal operation. In this paper a novel input vector monitoring concurrent BIST scheme is presented, that reduces considerably the imposed hardware overhead compared to previously proposed schemes.

[255] High Performance Reliable Variable Latency Carry Select Addition [p. 1257]
K Du, P Varman and K Mohanram
Speculative adders have attracted strong interest for reducing critical path delays to sub-logarithmic delays by exploiting the tradeoffs between reliability and performance. Speculative adders also find use in the design of reliable variable latency adders, which combine speculation with error correction to achieve high performance for low area overhead over traditional adders. This paper describes speculative carry select addition (SCSA), a novel function speculation technique for the design of low error-rate speculative adders and low overhead, high performance, reliable variable latency adders. We develop an analytical model for the error rate of SCSA to facilitate both design exploration and convergence. We show that for an error rate of 0.01% (0.25%), SCSA-based speculative addition is 10% faster than the DesignWare adder with up to 43% (56%) area reduction. Further, on average, variable latency addition using SCSA-based speculative adders is 10% faster than the DesignWare adder with area requirements of -19% to 16% (-17% to 29%) for unsigned random (signed Gaussian) inputs.

Defect density and variabilities in values of parameters continue to grow with each new generation of nano-scale fabrication technology. In SRAMs, variabilities reduce yield and necessitate extensive interventions, such as the use of increasing numbers of spares to achieve acceptable yield. For most microprocessor chips, the number of SRAM bits is expected to grow 2... for every generation. Consequently, microprocessor chip yields will be seriously undermined if no defect-tolerance approach is used. In this paper, we show the limits of the traditional spares-based defect-tolerance approaches for SRAMs. We then propose and implement a software-based approach for improving cache yield. We demonstrate that our approach can significantly increase microprocessor chip yields (normalized with respect to chip area) compared to the traditional approaches, for upcoming fabrication technologies. In particular, we demonstrate that our approach dramatically increases effective computing capacity, measured in MIPS-per-unit-chip-area. Our approach does not require any hardware design changes and hence can be applied to improve yield of any modern microprocessor chip, incurs low performance penalty only for the chips with unrepaired defects in SRAMs, and adapts without requiring any design changes as the yield improves for a particular design and fabrication technology.

Lifetime (long-term) reliability has been a main design challenge as technology scaling continues. Time-dependent dielectric breakdown (TDBB), negative bias temperature instability (NBTI), and electromigration (EM) are some of the critical failure mechanisms affecting lifetime reliability. Due to the correlation between different failure mechanisms and their significant dependence on the operating temperature, existing models assuming constant failure rate and additive impact of failure mechanisms will underestimate the lifetime of a system, usually measured by mean-time-to-failure (MTTF). In this paper, we propose a new methodology which evaluates system lifetime in MTTF and relies on Monte-Carlo simulation for verifying results. Temperature variations and the correlation between failure mechanisms are considered so as to mitigate lifetime underestimation. The proposed methodology, when applied on an Alpha 21264 processor, provides less pessimistic lifetime evaluation than the existing models based on sum of failure rate. Our experimental results also indicate that, by considering the correlation of TDBB and NBTI, the lifetime of a system is likely not dominated by TDBB or NBTI, but by EM or other failure mechanisms.

Moore's Law has driven the semiconductor revolution enabling over four decades of scaling in frequency, size, complexity, and power. However, the limits of physics are preventing further scaling of speed, forcing a paradigm shift towards multicore computing and parallelization. In effect, the system is taking over the role that the single CPU was playing: high-speed signals running through chips but also packages and boards connect ever more complex systems. High-speed signals making their way through the entire system cause new challenges in the design of computing hardware. Inductance, phase shifts and velocity of light effects, material resonances, and wave behavior become not only prevalent but need to be calculated accurately and rapidly to enable short design cycle times. In essence, to continue scaling with Moore's Law requires the incorporation of Maxwell's equations in the design process. Incorporating Maxwell's equations into the design flow is only possible through the combined power that new algorithms, parallelization and high-speed computing provide. At the same time, incorporation of Maxwell-based models into circuit and system-level simulation presents a massive accuracy, passivity, and scalability challenge. In this tutorial, we navigate through the often confusing terminology and concepts behind field solvers, show how advances in field solvers enable integration into EDA flows, present novel methods for model generation and passivity assurance in large systems, and demonstrate the power of cloud computing in enabling the next generation of scalable Maxwell solvers and the next generation of scalable Maxwell solvers and the next generation.
11.1: SPECIAL DAY MORE-THAN-MOORE: Heterogeneous Integration

Moderator: M Brillouët, CEA-Leti, FR

Challenges and Emerging Solutions in Testing TSV-Based 2.5D-and 3D-Stacked ICs - Invited Paper [p. 1277]
E J Marinissen

Through-Silicon Vias (TSVs) provide high-density, low-latency, and low-power vertical interconnects through a thinned-down wafer substrate, thereby enabling the creation of 2.5D- and 3D-Stacked ICs. In 2.5D-SICs, multiple dies are stacked side-by-side on top of a passive silicon interposer base containing TSVs. 3D-SICs are towers of vertically stacked active dies, in which the vertical inter-die interconnects contain TSVs. Both 2.5D- and 3D-SICs are fraught with test challenges, for which solutions are only emerging. In this paper, we classify the test challenges as (1) test flows, (2) test contents, and (3) test access.

11.2: The Quest for NoC Performance

Moderators: D Bertozzi, Ferrara U, IT; C Seiculescu, EPF Lausanne, CH

A TDM NoC Supporting QoS, Multicast, and Fast Connection Set-Up [p. 1283]
R Stefan, A Molnos, A Ambrose and K Goossens

Networks-on-Chip are seen as promising interconnect solutions, offering the advantages of scalability and high frequency operation which the traditional bus interconnects lack. Several NoC implementations have been presented in the literature, some of them having mature tool-flows and ecosystems. The main differentiating factor between the various implementations are the services and communication patterns they offer to the enduser. In this paper we present dAEelite, a TDM Network-on-Chip that offers a unique combination of features, namely guaranteed bandwidth and latency per connection, built-in support for multicast, and a short connection set-up time. While our NoC was designed from the ground up, we leverage on existing tools for network dimensioning, analysis and instantiation. We have implemented and tested our proposal in hardware and we found it to compare favorably to the other NoCs in terms of hardware area. Compared with aelite, which is closest in terms of offered services our network offers connection set-up times faster by a factor of 10 network, traversal latencies decreased by 33%, and improved bandwidth.

Parallel Probing: Dynamic and Constant Time Setup Procedure in Circuit Switching NoC [p. 1289]
S Liu, A Jantsch and Z Lu

We propose a circuit switching Network-on-chip with a parallel probe searching setup method, which can search the entire network in constant time, only dependent on the network size but independent of the network load. Under a specific search policy, the setup procedure is guaranteed to terminate in time 3D+6 cycles, where D is the geometric distance between source and destination. If a path can be found, the method succeeds in 3D+6 cycles, if a path cannot be found, it fails in maximum 3D+6 cycles. Compared to previous work, our method can reduce the setup time and enhance the success rate of setups. Our experiments show that compared with a sequential probe searching method, this method can reduce the search time by up to 20%. Compared with a centralized channel allocator method, this method can enhance the success rate by up to 20%.

A Flit-level Speedup Scheme for Network-on-Chips Using Self-Reconfigurable Bi-directional Channels [p. 1295]
Z Qian, Y F Teh and C-Y Tsui

In this work, we propose a flit-level speedup scheme to enhance the network-on-chip (NoC) performance utilizing bidirectional channels. In addition to the traditional efforts on allowing flits of different packets using the idling internal and external bandwidth of the bi-directional channel, our proposed flit-level speedup scheme also allows flits within the same packet to be transmitted simultaneously on the bi-directional channel. For inter-router transmission, a novel distributed channel configuration protocol is developed to dynamically control the link directions. For the intra-router transmission, an input buffer architecture which supports reading and writing two flits from the same virtual channel at the same time is proposed. The switch allocator is also designed to support flit-level parallel arbitration. Simulation results on both synthetic traffic and real benchmarks show performance improvement in throughput and latency over the existing architectures using bi-directional channels.

Index Terms - Bidirectional channel, flit-level speedup, NoC
As the technology scales down, the increased power density brings in significant system reliability issues. Therefore, the temperature monitoring and the induced power management become more and more critical. The thermal fluctuation effects of the recently discovered spintronic memristor make it a promising candidate as a temperature sensing device. In this paper, we carefully analyzed the thermal fluctuations of spintronic memristor and the corresponding design considerations. On top of it, we proposed a temperature sensing circuit design by combining spintronic memristor with the traditional CMOS current reference. Our simulation results show that the proposed design can provide high accuracy of temperature detection within much smaller footprint compared to the traditional CMOS temperature sensor designs. As magnetic device scales down, the relatively high power consumption is expected to be reduced.

Keywords: thermal sensor; memristor; current reference; temperature

Three-dimensional (3D) flash memory is emerging to fulfill the ever-increasing demands of storage capacity. In 3D NAND flash memory, multiple layers are stacked to increase bit density and reduce bit cost of flash memory. However, the physical architecture of 3D flash memory leads to a higher probability of disturbance to adjacent physical pages and greatly increases bit error rates. This paper presents 3D-FlashMap, a novel physical-location-aware block mapping strategy for three-dimensional NAND flash memory. 3D-FlashMap permutes the physical mapping of blocks and maximizes the distance between consecutively logical blocks, which can significantly reduce the disturbance to adjacent physical pages and effectively enhance the reliability. We apply 3D-FlashMap to a representative flash storage system. Experimental results show that the proposed scheme can reduce uncorrectable page errors by 85% with less than 2% space overhead in comparison with the baseline scheme.

As one promising candidate for next-generation nonvolatile memory technologies, spin-transfer torque random access memory (STT-RAM) has demonstrated many attractive features, such as nanosecond access time, high integration density, non-volatility, and good CMOS process compatibility. In this paper, we reveal an important fact that has been neglected in STT-RAM designs for long: the write operation of a STT-RAM cell is asymmetric based on the switching direction of the MTJ (magnetic tunneling junction) device: the mean and the deviation of the write latency for the switching from low- to high-resistance state is much longer or larger than that of the opposite switching. Some special design concerns, e.g., the write-pattern-dependent write reliability, are raised by this observation. We systematically analyze the root reasons to form the asymmetric switching of the MTJ and study their impacts on STT-RAM write operations. These factors include the thermal-induced statistical MTJ magnetization process, asymmetric biasing conditions of NMOS transistors, and both NMOS and MTJ device variations. We also explore the design space of different design methodologies on capturing the switching asymmetry of different STT-RAM cell structures. Our experiment results proved the importance of full statistical design method in STT-RAM designs for design pessimism minimization.

In this publication we present the results of our investigations into the reliability and uniqueness of Static Random Access Memories (SRAMs) in different technology nodes when used as a Physically Unclonable Function (PUF). The comparative analysis that can be found in this publication is the first ever of its kind, using different SRAM memories in technologies ranging from 180nm to 65nm. Each SRAM memory presents a unique and unpredictable start-up pattern when being powered up. In order to use an SRAM as a PUF in an application, the stability of its start-up patterns needs to be assured under a wide variety of conditions such as temperature and applied voltage. Furthermore the start-up patterns of different memories must be unique and contain sufficient entropy. This paper presents the results of tests that investigate these properties of different SRAM memory technology nodes. Furthermore, it proposes the construction of a fuzzy extractor, which can be used in combination with the tested memories for extracting secure cryptographic keys.
Many True Random Numbers Generators (TRNG) use jittery clocks generated in ring oscillators as a source of entropy. This is especially the case in Field Programmable Gate Arrays (FPGA), where sources of randomness are very limited. Inverter Ring Oscillators (IRO) are relatively well characterized as entropy sources. However, it is known that they are very sensitive to working conditions. This fact makes them vulnerable to attacks. On the other hand, Self-Timed Rings (STR) are currently considered as a promising solution to generate robust clock signals. Although many studies deal with their temporal behavior and robustness in Application Specific Integrated Circuits (ASIC), equivalent study does not exist for FPGAs. Furthermore, these oscillators were not analyzed and characterized as entropy sources aimed at TRNG design. In this paper, we analyze STRs as entropy sources for TRNGs implemented in FPGAs. Next, we compare STRs and IROs when serving as sources of randomness. We show that STRs represent very interesting alternative to IROs: they are more robust to environmental fluctuations and they exhibit lower extra-device frequency variations.

This work offers a framework which does not rely on a Golden IC (GIC) during hardware Trojan (HT) detection. GIC is a Trojan-free IC which is required, in all existing HT frameworks, as a reference point to verify the responses obtained from an IC under authentication. However, identifying a GIC is not a trivial task. A GIC may not even exist, since all the fabricated ICs may be HT-infected. We propose a framework which is based on adding a set of detection sensors to a design which are integrated in the free spaces on the layout and fabricated on the same die. After fabrication, a self-authentication procedure is proposed in order to determine if a Trojan is inserted in a set of arbitrarily-selected paths in the design. The detection process uses on-chip measurements on the sensors and the design paths in order to evaluate the correlation between a set of actual and predicted delay ranges. Error in the on-chip measurement infrastructure is considered. If our framework determines that a Trojan is (or is not) inserted on a considered path, then it is accurate. In our computational experiments, conducted for challenging cases of small Trojan circuits in the presence of die-to-die and within-die process variations, we report a high detection rate to show its effectiveness in realizing a self-authentication process which is independent of a GIC.

Large analog circuit models are very expensive to evaluate and verify. New techniques are needed to shorten time-to-market and to reduce the cost of producing a correct analog integrated circuit. Model order reduction is an approach used to reduce the computational complexity of the mathematical model of a dynamical system, while capturing its main features. This technique can be used to reduce an analog circuit model while retaining its realistic behavior. In this paper, we present an approach to model order reduction of nonlinear analog circuits. We model the circuit using fuzzy differential equations and use qualitative simulation and K-means clustering to discretize efficiently its state space. Moreover, we use a conformance checking approach to refine model order reduction steps and guarantee simulation acceleration and accuracy. In order to illustrate the effectiveness of our method, we applied it to a transmission line with nonlinear diodes and a large nonlinear space. Furthermore, we use a conformance checking approach to refine model order reduction steps and guarantee simulation acceleration and accuracy. In order to illustrate the effectiveness of our method, we applied it to a transmission line with nonlinear diodes and a large nonlinear space. Experimental results show that our reduced models are more than one order of magnitude faster and accurate when compared to existing methods.

The efficiency of different assist techniques for SRAM cell functionality improvement under the influence of random process variation is studied in this paper. The sensitivity of an SRAM cell functionality metrics when using control voltage level assist techniques is analyzed in read and write operation modes. The efficiency of the assist techniques is estimated by means of parametric analysis. The purpose is to find the degree of functionality metric improvement in each operation mode. The Acceptance Region concept is used for parametric analysis of SRAM cell functionality under random threshold voltage variations. In order to increase the reliability of the SRAM several assist techniques, chosen among the most efficient ones for each operation mode, are considered. This analysis offers a qualitative indication of the cell's functionality improvement by means of the efficient computation of a metric in parameter domain analysis. The results are proven to have high correlation with the ones obtained by means of the classical Monte Carlo simulations with significant savings in comparing different assist techniques.

Keywords: SRAM Cell; Voltage Level Assist Techniques; Spec Violation Metric;
In this paper, we propose a new envelope-following parallel transient analysis method for the general switching power converters. The new method first exploits the parallelism in the envelope-following method and parallelize the Newton update solving part, which is the most computational expensive, in GPU platforms to boost the simulation performance. To further speed up the iterative GMRES solving for Newton update equation in the envelope-following method, we apply the matrix-free Krylov basis generation technique, which was previously used for RF simulation. Last, the new method also applies more robust Gear-2 integration to compute the sensitivity matrix instead of traditional integration methods. Experimental results from several integrated on-chip power converters show that the proposed GPU envelope-following algorithm leads to about 10x speedup compared to its CPU counterpart, and 100x faster than the traditional envelope-following methods while still keeping the similar accuracy.

The numerical calculation of the limit cycle of oscillators with resonators exhibiting a high quality factor such as quartz crystals is a difficult task in the time domain. Time domain integration formulas introduce numerical damping which leads asymptotically to erroneous limit cycles or spurious oscillations. The numerical problems for solving the underlying differential algebraic equations are studied in detail. Based on these results a class of novel integration formulas is derived and the results are compared with the well-known Harmonic Balance (HB) technique. The discretized system is sparser than that of the HB method, therefore easier to solve and slower run time.

NANO- ELECTRO-MECHANICAL RELAYS FOR FPGA ROUTING: EXPERIMENTAL DEMONSTRATION AND A DESIGN TECHNIQUE


Nanoelectro-Mechanical (NEM) relays are excellent candidates for programmable routing in Field Programmable Gate Arrays (FPGAs). FPGAs that combine CMOS circuits with NEM relays are referred to as CMOS-NEM FPGAs. In this paper, we experimentally demonstrate, for the first time, correct functional operation of NEM relays as programmable routing switches in FPGAs, and their programmability by utilizing hysteresis properties of NEM relays. In addition, we present a technique that utilizes electrical properties of NEM relays and selectively removes or downsizes routing buffers for designing energy-efficient CMOS-NEM FPGAs. Simulation results indicates that such CMOS-NEM FPGAs can achieve 10-fold reduction in leakage power, 2-fold reduction in dynamic power, and 2-fold reduction in area, simultaneously, without application speed penalty when compared to a 22nm CMOS-only FPGA.

State-based Full Predication for Low Power Coarse-Grained Reconfigurable Architecture

K Han, S Park and K Choi

It has been one of the most fundamental challenges in architecture design to achieve high performance with low power while maintaining flexibility. Parallel architectures such as coarse-grained reconfigurable architecture, where multiple PEs are tightly coupled with each other, can be a viable solution to the problem. However, the PEs are typically controlled by a centralized control unit, which makes it hard to parallelize programs requiring different control of each PE. To overcome this limitation, it is essential to convert control flows into data flows by adopting the predicated execution technique, but may incur additional power consumption. This paper reveals power issues in the predicated execution and proposes a novel technique to mitigate power overhead of predicated execution. Contrary to the conventional approach, the proposed mechanism can decide whether to suppress instruction execution or not without decoding the instructions and does not require additional instruction bits, thereby resulting in energy savings. Experimental results show that energy consumed by the reconfigurable array and its configuration memory is reduced by up to 23.9%.

UPaRC -- Ultra-Fast Power-aware Reconfiguration Controller

R Bonamy, H-M Pham, S Pillement and D Chillet

Dynamically reconfigurable architectures, which can offer high performance, are increasingly used in different domains. High-speed reconfiguration process can be carried out by operating at high frequency but can also augment the power consumption. Thus the effort on increasing performance by accelerating the reconfiguration should take into account power consumption constraints. In this paper, we present...
an ultra-fast power-aware reconfiguration controller (UPaRC) to boost the reconfiguration throughput up to 1.433 GB/s. UPaRC can not only enhance the system performance, but also auto-adapt to various performance and consumption conditions. This could enlarge the range of applications and optimize for each selected application during run-time. An investigation of reconfiguration bandwidths at different frequencies and with different bitstream sizes are experimentally quantified and presented. The power consumption measurements are also realized to emphasize energy-efficiency of UPaRC over state-of-the-art reconfiguration controllers up to 45 times more efficient.

Index Terms - dynamic partial reconfiguration, rapid reconfiguration speed, power consumption, ICAP

G Mariani, V-M Sirna, G Palemo, V Zaccara, C Silvano and K Bertels

Resource run-time managers have been shown particularly effective for coordinating the usage of the hardware resources by multiple applications, eliminating the necessity of a full-blown operating system. For this reason, we expect that this technology will be increasingly adopted in emerging multi-application reconfigurable systems. This paper introduces a fully automated design flow that exploits multi-objective design space exploration to enable run-time resource management for the Molen reconfigurable architecture. The entry point of the design flow is the application source code; our flow is able to heuristically determine a set of candidate hardware/software configurations of the application (i.e., operating points) that trade off the occupation of the reconfigurable fabric (in this case, an FPGA), the load of the master processor and the performance of the application itself. This information enables a run-time manager to exploit more efficiently the available system resources in the context of multiple applications. We present the results of an experimental campaign where we applied the proposed design flow to two reference audio applications mapped on the Molen architecture. The analysis proved that the overhead of the design space exploration and operating points extraction with respect to the original Molen flow is within reasonable bounds since the final synthesis time still represents the major contribution. Besides, we have found that there is a high variance in terms of execution time speedup associated with the operating points of the application (characterized by a different usage of the FPGA) which can be exploited by the run-time manager to increase/decrease the quality of service of the application depending on the available resources.

11.7: Rise and Fall of Layout
Moderators: R Otten, TU Eindhoven, NL; P Groeneveld, Magma Design Automation, US

[277] VLSI Legalization with Minimum Perturbation by Iterative Augmentation [p. 1385]
U Brenner

We present a new approach to VLSI placement legalization. Based on a minimum-cost flow algorithm that iteratively augments flows along paths, our algorithm ensures that only augmentations are considered that can be realized exactly by cell movements. Hence, the method avoids realization problems which are inherent to previous flow-based legalization algorithms. As a result, it combines the global perspective of minimum-cost flow approaches with the efficiency of local search algorithms. The tool is mainly designed to minimize total and maximum cell movement but it is flexible enough to optimize the effect on timing or netlength, too. We compare our approach to legalization tools from industry and academia by experiments on dense recent real-world designs and public benchmarks. The results show that we are much faster and produce significantly better results in terms of average (linear and quadratic) and maximum movement than any other tool.

[278] Agglomerative-Based Flip-Flop Merging with Signal Wirelength Optimization [p. 1391]
S S-Y Liu, C-J Lee and H-M Chen

In this paper, an optimization methodology using agglomerative-based clustering for number of flip-flop reduction and signal wirelength minimization is proposed. Comparing to previous works on flip-flop reduction, our method can obtain an optimal tradeoff curve between flip-flop number reduction and increase in signal wirelength. Our proposed methodology outperforms [1] and [12] in both reducing number of flip-flops and minimizing increase in signal wirelength. In comparison with [9], our methodology obtains a tradeoff of 15.8% reduction in flip-flop’s signal wirelength with 16.9% additional flip-flops. Due to the nature of agglomerative clustering, when relocating flipflops, our proposed method minimizes total displacement by an average of 5.9%, 8.0%, 181.4% in comparison with [12], [1] and [9] respectively.

[279] Fixed Origin Corner Square Inspection Layout Regularity Metric [p. 1397]
M Pons, M Morgan and C Piguet

Integrated circuits suffer from serious layout printability issues associated to the lithography manufacturing process. Regular layout designs are emerging as alternative solutions to help reducing these systematic subwavelength lithography variations. However, there is no metric to evaluate and compare the layout regularity of those regular designs and there is no methodology to link layout regularity to the reduction of process variations. In this paper we propose a new layout regularity metric called Fixed Origin Corner Square Inspection (FOCSI). We also provide a methodology using the Monte Carlo analysis to evaluate and understand the impact of regularity on process variability.
We discuss three complementary approaches that can provide both portability and an increased level of abstraction for the programming of heterogeneous multicore systems. Together, these approaches also support performance portability, as currently investigated in the EU FP7 project PEPPHER. In particular, we consider (1) a library-based approach, here represented by the integration of the SkePU C++ skeleton programming library with the StarPU runtime system for dynamic scheduling and dynamic selection of suitable execution units for parallel tasks; (2) a language-based approach, here represented by the Offload-C++ high-level language extensions and Offload compiler to generate platform-specific code; and (3) a component-based approach, specifically the PEPPHER component system for annotating user-level application components with performance metadata, thereby preparing them for performance-aware composition. We discuss the strengths and weaknesses of these approaches and show how they could complement each other in an integrational programming framework for heterogeneous multicore systems.

In this paper, we present a variational electromagnetic-semiconductor coupled solver to assess the impacts of process variations on the 3D integrated circuit (3D IC) on-chip structures. The solver employs the finite volume method (FVM) to handle a system of equation considering both the full-wave electromagnetic effects and semiconductor effects. With a smart geometrical variation model for the FVM discretization, the solver is able to handle both small-size or large-size variations. Moreover, a weighted principle factor analysis (wPFA) technique is presented to reduce the random variables in both electromagnetic and semiconductor regions, and the spectral stochastic collocation method (SSCM) [10] is used to generate the quadratic statistical model. Numerical results validate the accuracy and efficiency of this solver in dealing with process variations in hybrid material through-silicon via (TSV) structures.

We present a variant of dynamic time warping (DTW) algorithm to verify jitter properties associated with an analog and mixed signal (AMS) design. First, the AMS design with stochastic jitter component is modeled using a system of difference equations for analog and digital parts and then evaluated in a MATLAB simulation environment. Second, MonteCarlo simulation is combined with DTW and hypothesis testing to determine the probability of acceptance/rejection of those simulation results. Our approach is illustrated on analyzing the jitter effect on the "lock-time" property of a phase locked loop (PLL) based frequency synthesizer.

Cyber-physical systems have become more difficult to test as hardware and software complexity grows. The increased integration between computing devices and physical phenomena demands new techniques for ensuring correct operation of devices across a broad range of operating conditions. Manual test methods, which involve test personnel, require much effort and expense and lengthen a device's time to market. We describe a method for test automation of devices wherein a device is connected to a digital mockup of the physical environment, where both the device and the digital mockup are managed by PC-based software. A digital mockup consists of a behavioral model of the interacting environment, such as a medical ventilator device connected to a digital mockup of human lungs. We introduce Mockup Electronic Data Sheets (MEDS) as a method for embedding model information into the digital mockup, allowing PC software to automatically detect configurable model parameters and facilitate test automation. We summarize a case study showing the effectiveness of digital mockups and MEDS as a framework for test automation on a medical ventilator, resulting in 5x less time spent testing compared to methods requiring test personnel.
This paper presents component-based and aspect-oriented methodology and tool for designing and developing Real-Time Embedded Control Systems (RTECS). This methodology defines a component model for describing modular and reusable software to cope with the increasing complexity of embedded systems. It proposes an aspect-oriented approach to address explicitly the extra-functional concerns of RTECS, to describe separately transversal real time and security constraints, and to support model properties analysis. The benefits of this methodology are shown via an example of Legway control software, a version of the Segway vehicle built with Lego Mindstorms NXT.

Keywords - Model-based design, software component, aspect-oriented programming, embedded system design, embedded control software.
A Block-Level Flash Memory Management Scheme for Reducing Write Activities in PCM-based Embedded Systems [p. 1447]

D Liu, T Wang, Y Wang, Z Qin and Z Shao

This paper targets at an embedded system with phase change memory (PCM) and NAND flash memory. Although PCM is a promising main memory alternative and is recently introduced to embedded system designs, its endurance keeps drifting down and greatly limits the lifetime of the whole system. Therefore, this paper presents a block-level flash memory management scheme, WAB-FTL, to effectively manage NAND flash memory while reducing write activities of the PCM-based embedded systems. The basic idea is to preserve each bit in flash mapping table hosted by PCM from being inverted frequently during the process of mapping table update. To achieve this, a new merge strategy is adopted in WAB-FTL to delay the mapping table update, and a tiny mapping buffer is used for caching frequently updated mapping records. Experimental results based on Android traces show that WAB-FTL can effectively reduce write activities when compared with the baseline scheme.

Architecting a Common-Source-Line Array for Bipolar Non-Volatile Memory Devices [p. 1451]

B Zhao, J Yang, Y Zhang, Y Chen and H Li

Traditional array organization of bipolar nonvolatile memories such as STT-MRAM and memristor utilizes two bitlines for cell manipulations. With technology scaling, such bitline pair will soon become the bottleneck of density improvement. In this paper we propose a novel common-source-line array architecture, which uses a shared source-line along the row, leaving only one bitline per column. We also elaborate our design flow towards a reliable common-source-line array design, and demonstrate its effectiveness on STT-MRAM and memristor memory arrays. Our study results show that with comparable latency and energy, the proposed common-source-line array can save 33% and 21.8% area for Memristor-RAM and STT-MRAM respectively, comparing with corresponding traditional dual-bitline array designs.

Layout-Aware Optimization of STT MRAMs [p. 1455]

S K Gupta, S P Park, N N Mojumder and K Roy

We present a layout-aware optimization methodology for spin-transfer torque (STT) MRAMs, considering the dependence of cell area on the access transistor width (WFET), number of fingers in the access transistor and the metal pitch of bit- and source-lines. It is shown that for WFET less than a critical value (~7 times the minimum feature length), one-finger transistor yields minimum cell area. For large WFET, minimum cell area is achieved with a two-finger transistor. We also show that for a range of WFET, the cell area is limited by the metal pitch of bit- and source-lines. A result, in the metal pitch limited (MPL) region, WFET can be increased with no change in the cell area. We analyze the impact of increase in WFET in the MPL region on the write margin and cell tunneling magneto-resistance (CTMR) of different genres of STT MRAMs. We consider conventional STT MRAM cells in the standard and reverse-connected configurations and STT MRAMs with tilted magnetic anisotropy for the analysis. By increasing WFET from the minimum to the maximum value in the MPL region (at iso-cell area) and reducing read voltage to achieve iso-read disturb margin, 2X improvement in write margin and 27% improvement in CTMR is achieved for the reverse-connected STT MRAM. Similar trends are observed for other STT MRAM cells.

Characterization of the Bistable Ring PUF [p. 1459]

Q Chen, G Csaba, P Lugli, U Schlichtmann and U Ruehrmair

The bistable ring physical(y) unclonable function (BRPUF) is a novel electrical intrinsic PUF design for physical cryptography. FPGA prototyping has provided a proof-of-concept, showing that the BR-PUF could be a promising candidate for strong PUFs. However, due to the limitations (device resources, placement and routing) of FPGA prototyping, the effectiveness of a practical ASIC implementation of the BRPUF could not be validated. This paper characterizes the BRPUF further through transistor-level simulations. Based on process variation, mismatch, and noise models provided or suggested by industry, these simulations are able to provide predictions on the figures-of-merit of ASIC implementations of the BR-PUF. This paper also suggests a more secure way of using the BR-PUF based on its supply voltage sensitivity.

An Operational Matrix-Based Algorithm for Simulating Linear and Fractional Differential Circuits [p. 1463]

Y Wang, H Liu, G K H Pang and N Wong

We present a new time-domain simulation algorithm (named OPM) based on operational matrices, which naturally handles system models cast in ordinary differential equations (ODEs), differential algebraic equations (DAEs), high-order differential equations and fractional differential equations (FDEs). When applied to simulating linear systems (represented by ODEs or DAEs), OPM has similar performance to
advanced transient analysis methods such as trapezoidal or Gear's method in terms of complexity and accuracy. On the other hand, OPM naturally handles FDEs without much extra effort, which can not be efficiently solved using existing time-domain methods. High-order differential systems, being special cases of FDEs, can also be simulated using OPM. Moreover, adaptive time step can be utilized in OPM to provide a more flexible simulation with low CPU time. Numerical results then validate OPM's wide applicability and superiority.

T Cupaiuolo and D Lo Iacono

The importance of having an efficient Fast Fourier Transform (FFT) implementation is universally recognized as one of the key enablers for the development of new and more powerful signal processing algorithms. In the field of telecommunications, one of its most recent applications is the Orthogonal Frequency Division Multiplexing (OFDM) modulation technique, whose superiority is recognized and endorsed by several standards. However, the horizon of standards is so wide and heterogeneous that a single FFT implementation hardly satisfies them all. In order to have a reusable, easily extensible and reconfigurable solution, most of the baseband processing is moving towards a software implementation: to this end several new Digital Signal Processor (DSP) architectures are emerging, each with its own set of differentiating properties. Within this context, we propose a software implementation of the FFT on the Block Processing Engine (BPE) platform. Several implementations have been investigated, ranging from a single instruction based approach, to others employing several instructions either in parallel or in pipeline. The outcome is a flexible set of solutions that leaves degrees of freedom in terms of computational load, achievable throughput and power consumption. The proposed implementations closely approach the theoretical clock cycles expected by dedicated hardware counterpart, thus making it a concrete alternative.

M Mittag, A Krinke, G Jerke and W Rosenstiel

In industrial environments, full-custom layout design of analog and mixed-signal ICs is done hierarchically. In order to increase design efficiency, cell layouts are reused in the design hierarchy. Constraints forming relations between instances in different hierarchical contexts are of critical importance. While implementing a cell layout, these constraints have to be available in the cell’s context. In this paper, a general definition of hierarchical constraints for a constraint-driven design flow is given. Furthermore it is shown, how top-down declared constraints can be propagated into another hierarchical context. Only by propagation they become visible and verifiable for bottom-up cell design. The feasibility of our proposed methodology is shown by applying it to a modular Smart Power IC of the automotive industry.

I S Abed and A G Wassal

Double patterning lithography (DPL) is seen as one of the most promising solutions for new technology nodes such as 32nm and 22nm. However, DPL faces the challenges of handling layout decomposition and overlay errors. Currently, most DPL solutions use post-layout decomposition which requires multiple iterations and designer intervention to achieve a decomposable layout as designs scale larger. Recent research is starting to consider DPL constraints during the layout design phase especially during the detailed routing phase. In this work, we propose DPL-aware grid-based detailed routing algorithm supported with online conflict resolution. The conflict resolution algorithm uses a graph structure to represent geometrical relations between routed polygons and helps in conflict detection and color assignment. Experimental results indicate that this enhanced algorithm reduces the number of conflicts by 60% on average.

H-P Tsai, R-B Lin and L-C Lai

This paper presents a simple method for design and analysis of a via-configurable routing fabric formed by an array of routing fabric blocks (RFBs). The method simply probes into an RFB rather than resorts to full-chip routing to collect some statistics for a metric used to qualify the RFB. We find that the trade-off between wire length and via count is a good metric. This metric has been validated by full-chip routing and used successfully to create better routing fabrics.

12.1: SPECIAL DAY MORE-THAN-MOORE: Applications

Moderator: M Brilouët, CEA-Leti, FR

Towards A Wireless Medic Smart Card - Invited Paper [p. 1483]
Wireless data transmission has become an integral part of modern society and plays an increasingly important role in health care. Technology scaling is continuously increasing wireless data rates, thus allowing for more flexible high-speed interfaces, e.g., between medical imaging equipment and mass storage devices. However, one issue remains: The power consumption of high-speed wireless transceivers and non-volatile memory grows with the data rate. This prevents from innovations using these high-speed wireless interfaces in ultra-low power (or even energy-passive) medical equipment that can be used by patients without a heavy power source. Clear efforts are required to close this gap, i.e., to provide high-speed wireless solutions with reduced energy consumption per transmitted bit. As a very example, this work presents the concept of a wireless medical smart card that combines near field communication for authentication and low-speed signaling together with a 60GHz interface for fast wireless memory access in a single patient-owned ID card. The basic architecture, functionality and prospects of the concept are discussed. A power budget is calculated based on state-of-the-art technologies. To put the concept into practice, some necessary developments for a reduction of the power consumption are outlined.

A Fast, Source-Synchronous Ring-based Network-on-Chip Design [p. 1489]
A Mandal, S P Khatri and R N Mahapatra

Most network-on-chip (NoC) architectures are based on a mesh-based interconnection structure. In this paper, we present a new NoC architecture, which relies on source synchronous data transfer over a ring. The source synchronous ring data is clocked by a resonant clock, which operates significantly faster than individual processors that are served by the ring. This allows us to significantly improve the cross section bandwidth and the latency of the NoC. We have validated the design using a 22nm predictive process. Compared to the state-of-the-art mesh based NoC, our scheme achieves a 4.5x better bandwidth, 7.4x better contention free latency with 11% lower area and 35% lower power.

Area Efficient Asynchronous SDM Routers Using 2-Stage Clos Switches [p. 1495]
W Song, D Edwards, J Ganside and W J Bainbridge

Asynchronous on-chip networks are good candidates for multi-core applications requiring low-power consumption. Asynchronous spatial division multiplexing (SDM) routers provide better throughput with lower area overhead than asynchronous virtual channel routers; however, the area overhead of SDM routers is still significant due to their high-radix central switches. A new 2-stage Clos switch is proposed to reduce the area overhead of asynchronous SDM routers. It is shown that replacing the crossbars with the 2-stage Clos switches can significantly reduce the area overhead of SDM routers when more than two virtual circuits are used. The saturation throughput is slightly reduced but the area to throughput efficiency is improved. Using Clos switches increases the energy consumption of switches but the energy of buffers is reduced.

Power-Efficient Calibration and Reconfiguration for On-Chip Optical Communication [p. 1501]

On-chip optical communication infrastructure has been proposed to provide higher bandwidth and lower power consumption for the next-generation high performance multi-core systems. Online calibration of these optical components is essential to building a robust optical communication system, which is highly sensitive to process and thermal variation. However, the power consumption of existing tuning methods to properly calibrate the optical devices would be prohibitively high. We propose two calibration and reconfiguration techniques that can significantly reduce the worst case tuning power of a ring-resonator-based optical modulator: 1) a channel re-mapping scheme, with sub-channel redundant resonators, which results in significant reduction in the amount of required tuning, typically within the capability of voltage based tuning, and 2) a dynamic feedback calibration mechanism used to compensate for both process and thermal variations of the resonators. Simulation results demonstrate that these techniques can achieve a 48X reduction in tuning power - less than 10W for a network with 1-million ring resonators.

Modeling and Design Exploration of FBDRAM as On-chip Memory [p. 1507]
G Sun, C Xu and Y Xie

Compared to the traditional DRAM technology, floating body DRAM (FBDRAM) has many advantages, such as high density, fast access speed,
long retention time, etc. More important, FBDRAM is compatible with the traditional CMOS technology. It makes FBDRAM more competitive than other emerging memory technologies to be employed as on-chip memory. The characteristic variance of memory cells caused by process variations, however, has become an obstacle to adopt FBDRAM. In this work, we build a circuit level model of FBDRAM caches with corresponding architecture-level modifications to FBDRAM caches and study the trade-off among reliability, power consumption, and performance. With this model, we explore the L2 cache design using FBDRAM and compare it with traditional SRAM/eDRAM caches in both circuit and architectural levels.

Phase Change RAM (PCM) is a promising candidate of emerging memory technology to complement or replace existing DRAM and NAND Flash memory. A key drawback of PCMs is limited write endurance. To address this problem, several static wear-leveling methods that change logical to physical address mapping periodically have been proposed. Although these methods have low space overhead, they suffer from unnecessary data migrations thereby failing to exploit the full lifetime potential of PCMs. This paper proposes a new dynamic wear-leveling method that reduces unnecessary data migrations by adopting a hot/cold swapping-based dynamic method. Compared with the conventional hot/cold swapping-based dynamic method, the proposed method requires only a small amount of space overhead by applying Bloom filters to identify the hot and cold data. We simulate our method using SPEC2000 benchmark traces and compare with previous methods. Simulation results show that the proposed method reduces unnecessary data migrations by 58–92% and extends the memory lifetime by 2.18–2.30 times over previous methods with a negligible area overhead of 0.3%.

Nonvolatile processor has become an emerging topic in recent years due to its zero standby power, resilience to power failures and instant on feature. This paper first demonstrated a fabricated nonvolatile 8051-compatible processor design, which indicates the ferroelectric nonvolatile version leads to over 90% area overhead compared with the volatile design. Therefore, we proposed a compare and compress recovery architecture, consisting of a parallel run-length codec (PRLC) and a state table logic, to reduce the area of nonvolatile registers. Experimental results demonstrate that it can reduce the number of nonvolatile registers by 4 times with less than 1% overflow possibility, which leads to 43% overall processor area savings. Furthermore, we implemented the novel PRLC and defined the method to optimize the optimal parallel degree to accelerate the compressions. Finally, we proposed a reconfigurable state table architecture, which supports the reference vector selecting for different applications. With our heuristic vector selecting algorithm, the optimal vector can provide over 42% better register number reduction than other vector selecting approaches. Our method is also applicable to designs with other nonvolatile materials based registers.

12.4: Digital Communication Systems

Moderators: F Kienle, TU Kaiserslautern, DE; F Clemidy, CEA-Leti, FR

A Compression-based Area-efficient Recovery Architecture for Nonvolatile Processors [p. 1519]
Y Wang, Y Liu, Y Liu, D Zhang, S Li, B Sai, M-F Chiang and H Yang

Nonvolatile processor has become an emerging topic in recent years due to its zero standby power, resilience to power failures and instant on feature. This paper first demonstrated a fabricated nonvolatile 8051-compatible processor design, which indicates the ferroelectric nonvolatile version leads to over 90% area overhead compared with the volatile design. Therefore, we proposed a compare and compress recovery architecture, consisting of a parallel run-length codec (PRLC) and a state table logic, to reduce the area of nonvolatile registers. Experimental results demonstrate that it can reduce the number of nonvolatile registers by 4 times with less than 1% overflow possibility, which leads to 43% overall processor area savings. Furthermore, we implemented the novel PRLC and defined the method to optimize the optimal parallel degree to accelerate the compressions. Finally, we proposed a reconfigurable state table architecture, which supports the reference vector selecting for different applications. With our heuristic vector selecting algorithm, the optimal vector can provide over 42% better register number reduction than other vector selecting approaches. Our method is also applicable to designs with other nonvolatile materials based registers.

A Complexity Adaptive Channel Estimator for Low Power [p. 1531]
Z Yu, C H van Berkel and H Li

This paper presents a complexity adaptive channel estimator for low power. Channel estimation (CE) is one of the most computation intensive tasks in a software-defined radio (SDR) based OFDM demodulator. Complementary to the conventional low-power design methodology on processor architectures or circuits, we propose to reduce power also at the algorithm level. The idea is to dynamically scale the processing load of the channel estimator according to the run-time estimated channel quality. In this work, with a case study on China Mobile Multimedia Broadcasting (CMMB) standard, three practical CE algorithms are adopted to form a complexity scalable algorithm set, and signal noise ratio...
(SNR) is chosen to be the channel quality parameter for CE algorithm switching. In order to accurately estimate the SNR in the run-time, we also propose a noise variance estimation algorithm which is robust against fast-fading channels and introduces small computation overheads. Simulation shows that, under a pre-defined scenario for our targeting SDR demodulator, more than 50% run-time load reduction can be achieved compared with a fixed worst case channel estimator, while still fulfilling the mean square error requirement, resulting in about 25% of power reduction for the total demodulator. In addition, complexity adaption enables dynamical voltage and frequency scaling (DVFS) in a SDR demodulator which can lead to furthermore power reduction.

12.5: Architecture and Networks for Adaptive Computing

Moderators: F Ferrandi, Politecnico di Milano, IT; S Niar, Valenciennes U, FR


High performance hardware FFTs have numerous applications in instrumentation and communication systems. This paper describes a new parallel FFT architecture which combines the split-radix algorithm with a constant geometry interconnect structure. The split-radix algorithm is known to have lower multiplicative complexity than both radix-2 and radix-4 algorithms. However, it conventionally involves an "L-shaped" butterfly datapath whose irregular shape has uneven latencies and makes scheduling difficult. This work proposes a split-radix datapath that avoids the L-shape. With this, the split-radix algorithm can be mapped onto a constant geometry interconnect structure in which the wiring in each FFT stage is identical, resulting in low multiplexing overhead. Further, we exploit the lower arithmetic complexity of split-radix to lower dynamic power, by gating the multipliers during trivial multiplications. The proposed FFT achieves 46% lower power than a parallel radix-4 design at 4.5GS/s when computing a 128-point real-valued transform.


Hardware specialization is often the key to efficiency for programmable embedded systems, but comes at the expense of flexibility. This paper combines flexibility and efficiency in the design and synthesis of domain-specific datapaths. We merge all individual paths from the Data Flow Graphs (DFGs) of the target applications, leading to a minimal set of required resources; this set is organized into a column of physical operators and cloned, thus generating a domain-specific rectangular lattice. A bus-based FPGA-style interconnection network is then generated and dimensioned to meet the needs of the applications. Our results demonstrate that the lattice has good flexibility: DFGs that were not used as part of the datapath creation phase can be mapped onto it with high probability. Compared to an ASIC design of a single DFG, the speed of our domain-specific coarse-grained reconfigurable datapath is degraded by a factor up to 2x, compared to 3-4x for an FPGA; similarly, our lattice is up to 10x larger than an ASIC, compared to 20-40x for an FPGA. We estimate that our array is up to 6x larger than an ASIC accelerator, which is synthesized using datapath merging and has limited or null generality.


Embedded systems based on FPGA (Field-Programmable Gate Arrays) must exhibit more performance for new applications. However, no high-performance superscalar soft processor is available on the FPGA, because the superscalar architecture is not suitable for FPGAs. High-performance superscalar processors execute instructions out-of-order and it is necessary to re-order instructions after execution. This task is performed by the ROB (ReOrder Buffer) that uses usually multi-ports RAM, but only two-port buffers are available in FPGA. In this work, we propose a FPGA friendly ROB (ReOrder Buffer) architecture using only 2 ports RAM called a multi-bank ROB architecture. The ROB is the main and more complex structure in an out-of-order superscalar processor. Depending on processor architecture parameters, the FPGA implementation of our ROB compared to a classic architecture, requires 5 to 7 times less registers, 1.5 to 8.3 times less logic gates and 2.6 to 32 times less RAM blocks.

Keywords-component; superscalar processor; performance; out-of-order execution; FPGA; ReOrder Buffer


Processor architectures with Fine-Grained Reconfigurable Accelerators (FGRAs) allow for a high degree of adaptivity to address varying application requirements. When processing computation intensive kernels, multiple FGRAs may be used to execute a complex function. In order to exploit the adaptivity of a fine-grained reconfigurable fabric, a runtime system should decide when and which FGRAs to reconfigure with respect to application requirements. To enable this adaptivity, a flexible infrastructure is required that allows combining FGRAs to execute complex functions. We propose a mixed-grained reconfigurable architecture composed from a Coarse-Grained Reconfigurable Infrastructure (CGRI) that connects the FGRAs. At runtime we synthesize CGRI configurations that depend on decisions of the runtime system,
e.g. which FGRAs shall be reconfigured. Synthesis and place & route of the FGRAs are done at compile time for performance reasons. Combined, this results in a partial online synthesis for mixed-grained reconfigurable architectures, which allows maintaining a low runtime overhead while exploiting the inherent adaptivity of the reconfigurable fabric. In this work we focus on the crucial parts of synthesizing the configurations for the CGRI at runtime, propose algorithms, and compare their performance/overhead trade-offs for different application scenarios. We are the first to exploit the increased adaptivity of FGRAs that are connected by a CGRI, by using our partial online synthesis. In comparison to a state-of-the-art reconfigurable architecture that synthesizes the configurations for the CGRI at compile time we obtain an average speedup of 1.79x.

12.6: Boolean Methods in Logic Synthesis

Moderators: M Berkelaar, TU Delft, NL; J Monteiro, INESC-ID/TU Lisbon, PT

[312] Congestion-Aware Scheduling for NoC-based Reconfigurable Systems [p. 1561]
H-L Chao, Y-R Chen, S-Y Tung, P-A Hsiung and S-J Chen

Network-on-Chip (NoC) is becoming a promising communication architecture in place of dedicated interconnections and shared buses for embedded systems. Nevertheless, it has also created new design issues such as communication congestion and power consumption. A major factor leading to communication congestion is mapping of application tasks to NoC. Latency, throughput, and overall execution time are all affected by task mapping. As a solution, an efficient runtime Congestion-Aware Scheduling (CWS) is proposed for NoC-based reconfigurable systems, which predicts traffic pattern based on the link utilization. The proposed algorithm alleviates the overall congestion, instead of only improving the current packet blocking situation. Our experiment results have demonstrated that compared to other existing congestion-aware algorithms, the proposed CWS algorithm can reduce the average communication latency by 66%, increase the average throughput by 32%, reduce the energy consumption by 23%, and decrease the overall execution by 32%.

[313] Multi-Patch Generation for Multi-Error Logic Rectification by Interpolation with Cofactor Reduction [p. 1567]
K-F Tang, P-K Huang, C-N Chou and C-Y Huang

For a design with multiple functional errors, multiple patches are usually needed to correct the design. Previous works on logic rectification are limited to either single-fix or partial-fix rectifications. In other words, only one or part of the erroneous behaviors can be fixed in one iteration. As a result, it may lead to unnecessarily large patches or even failure in rectification. In this paper, we propose a multi-patch generation technique by interpolation with cofactor reduction. In particular, our method considers multiple errors in the design simultaneously and generates multiple patches to fix these errors. Experimental results show that the proposed method is effective on a set of large circuits, including the circuits synthesized from industrial register-transfer level (RTL) designs.

X Yang, T-K Lam, W-C Tang and Y-L Wu

Rewiring is a flexible and useful logic transformation technique through which a target wire can be removed by adding its alternative logics without changing the circuit functionality. In today's deep sub-micron era, circuit wires have become a dominating factor in most EDA processes and there are situations where removing a certain set of (perhaps extremely unwanted) wires is very useful. However, it has been experimentally suggested that the rewiring rate (percentage of original circuit wires being removable by rewiring) is only 30 to 40% for optimized circuits in the past. In this paper, we propose a generalized error cancellation modeling and flow to show that theoretically almost every circuit wire is removable under this flow. In the Flow graph Error Cancellation based Rewiring (FECR) scheme we propose here, a rewiring rate of 95% of even optimized circuits is obtainable under this scheme, affirming the basic claim of this paper. To our knowledge, this is the first known rewiring scheme being able to achieve this near complete rewiring rate. Consequently, this wire-removal process can now be considered as a powerful atomic and universal operation for logic transformations, as virtually every circuit node can also be removed through repetitions of this rewiring process. Besides, this modeling can also serve as a general framework containing many other rewiring techniques as its special cases.

S Ray, A Mishchenko, N Een, R Brayton, S Jang and C Chen

Mapping into K-input lookup tables (K-LUTs) is an important step in synthesis for Field-Programmable Gate Arrays (FPGAs). The traditional FPGA architecture assumes all interconnects between individual LUTs are "routable". This paper proposes a modified FPGA architecture which allows for direct (non-routable) connections between adjacent LUTs. As a result, delay can be reduced but area may increase. This paper investigates two types of LUT structures and the associated tradeoffs. A new mapping algorithm is developed to handle such structures. Experimental results indicate that even when regular LUT structures are used, area and delay can be improved 7.4% and 11.3%, respectively, compared to the high-effort technology mapping with structural choices. When the dedicated architecture is used, the delay can be improved
12.7: Impact of Modern Technology on Layout

Moderators: J Lienig, TU Dresden, DE; P Groeneveld, Magma Design Automation, US

[316] Row-Shift Decompositions for Index Generation Functions [p. 1585]
T Sasao

This paper shows a realization of incompletely specified index generation functions in the form \( f(x_1, x_2) = g(h(x_1) + x_2) \), where \(+\) denotes an integer addition. A decomposition algorithm is shown. Experimental results show that most of \( n = 2q-3 \) variable functions where \( k = 2^q - 1 \) combinations are specified can be realized by a pair of \( q \)-input \( q \)-output LUTs. The computation time is \( O(k) \). Experimental results using address tables, lists of English words, and randomly generated functions are shown.

Index Terms - Incompletely specified function, random function, linear transform, functional decomposition, data compression, IP address, hash function.

M Li, A Davoodi and L Xie

This work offers a framework for predicting the delays of individual design paths at the post-silicon stage which is applicable to post-silicon validation and delay characterization. The prediction challenge is mainly due to limited access for direct delay measurement on the design paths after fabrication, combined with the high degree of variability in the process and environmental factors. Our framework is based on using on-chip delay sensors to improve timing prediction. Given a placed netlist at the pre-silicon stage, an optimization procedure is described which automatically generates the sensors subject to an area budget and available whitespace on the layout, in the presence of process variations. Each sensor is then generated as a sequence of logic gates with an approximate location on the layout at the pre-silicon stage. The on-chip sensor delay is then measured to predict the delays of individual design paths with less pessimism. In our experiments, we show that custom on-chip sensors can significantly increase the rate of predicting if a specified set of paths are failing their timing requirements.

H-W Hsu, M-L Chen, H-M Chen, H-C Li and S-H Chen

Due to the advantage of flip-chip design in power distribution but controversial peripheral IO placement in lower design cost, redistribution layer (RDL) is usually used for such interconnection. Sometimes RDL is so congested that the capacity for routing is insufficient. Routing therefore cannot be completed within a single layer even for manual routing. Although [2] proposed a routing algorithm that uses two layers of RDLs, but in practice the required routing area is a little more than one layer. We overcome this problem by adopting the concept of pseudo single-layer. With the heuristics for routing on mapped channels and observations on staggered pins to relieve vertical constraints, the area of 2-layer routing can be minimized and the routability is 100%. Comparisons of routing results between manual design, the commercial tool, and the proposed method are presented. We have shown the effectiveness on a real industrial case: it originally required fully manual design, the proposed method can finish RDL routing automatically and effectively.

[319] AIR (Aerial Image Retargeting): A Novel Technique for In-Fab Automatic Model-Based Retargeting-for-Yield [p. 1603]
A Y Hamouda, M Anis and K S Karim

In this paper, we present a novel methodology for identifying lithography hot-spots and automatically transforming them into the lithography-friendly design space. This fast model-based technique is applied at the mask tape-out stage by slightly shifting and resizing the designs. It implicitly does a similar functionality as that of the Process Window OPC (PWOPC) but more efficiently. Being a relatively fast technique it also offers the means of providing the designer with all the design systematic deviations from the actual (on-wafer) parameters by including it in the parameter-extraction flow. We applied this methodology successfully to 28-nm Metal levels and showed that it efficiently (better quality and faster) improves the lithography-related yield and reliability issues.

Index Terms - RET, Resolution Enhancement Techniques, PWOPC, Optical Proximity Correction, LFD, lithography Friendly Design, DFM

[320] Layout-Driven Robustness Analysis for Misaligned Carbon Nanotubes in CNTFET-based Standard Cells [p. 1609]
M Beste and M B Tahoori

Carbon Nanotube Field Effect Transistors (CNTFETs) are being considered as a promising successor to current CMOS technology. Since the alignment of CNTs cannot be fully controlled yet, the layout of CNTFET-based standard cells has to be designed robust against misalignment. As CNTFET-based designs become more prevalent, a systematic methodology for misalignment robustness evaluation becomes crucial.

In this paper, we present a layout-driven robustness technique for misaligned carbon nanotubes in CNTFET-based standard cells. This method analyzes the impact of misalignment on the performance of standard cells and provides designers with a tool to optimize the layout for improved yield. The approach is based on a systematic analysis of the effect of misalignment on the electrical characteristics of CNTFETs, considering various types of misalignment. The methodology includes a simulation-based approach to evaluate the impact of misalignment on the performance of standard cells, enabling designers to make informed decisions during the layout design process.
this work we present a novel EDA tool "Layout-Driven Robustness Analysis" (LDRA) which enables designers to, for the first time, measure the robustness against misalignment. LDRA is validated and applied to various CNTFET-based standard cell layouts. The comparison of these cells reveals that robustness against misalignment is complex and depends on many factors. A CNT curve model is introduced and its influence on the robustness result is discussed. In conclusion, key factors for designing layouts robust against misalignment are proposed.

12.8: EMBEDDED TUTORIAL - Advances in Variation-Aware Modeling, Verification, and Testing of Analog ICs

Moderator: TBD

D De Jonghe, E Maricau, G Gielen, T McConaghy, B Tasic, and H Stratigopoulos

This tutorial paper describes novel scalable, nonlinear, generic, and industrially-oriented approaches to perform variation-aware modeling, verification, fault simulation, and testing of analog/custom ICs. In the first section, Dimitri De Jonghe, Elie Maricau, and Georges Gielen present a new advance extracting highly nonlinear, variation-aware behavioral models, through the use of data mining and a re-framing of the model-order reduction problem. In the next section, Trent McConaghy describes new statistical machine learning techniques that enable new classes of industrial EDA tools, which in turn are enabling designers to perform fast and accurate PVT / statistical / high-sigma design and verification. In the third section, Bratislav Tasic presents a novel industrially-oriented approach to analog fault simulation that also has applicability to variation-aware design. In the final section, Haralampos Stratigopoulos describes describes state-of-the-art analog testing approaches that address process varia

Source URL: http://www.date-conference.com/conference/proceedings/2012/abstracts

Links: