Architecture and IC Design, Embedded Software
CEA is a French government-funded technological research organization. Drawing on its excellence in fundamental research, its activities cover three main areas: Energy, Information and Health Technologies, and Defense and Security. As a prominent player in the European Research Area, with an internationally acknowledged level of expertise in its core competencies, CEA is involved in setting up collaborative projects with many partners around the world.

Within CEA Technological Research Division, three institutes lead researches in order to increase the industrial competitiveness through technological innovation and transfers: the CEA-LETI, focused on microelectronics, information & healthcare technologies, the CEA-LIST dedicated to technologies for digital systems, and the CEA-LITEN devoted to new energy technologies.

The CEA-LETI is focused on micro and nanotechnologies and their applications, from wireless devices and systems, to biology and healthcare or photonics. Nanoelectronics and Microsystems (MEMS) are at the core of its silicon activities. As a major player in the MINATEC innovation campus, CEA-LETI operates 8,000-m² state-of-the-art clean rooms, on 24/7 mode, on 200mm and 300mm wafer platforms. With 1,700 employees, CEA-LETI trains more than 240 Ph.D. students and hosts 200 assignees from partner companies. Strongly committed to the creation of value for the industry, CEA-LETI puts a strong emphasis on intellectual property and owns more than 1,880 patent families. For more information, visit http://www.leti.fr.

The CEA-LIST is a key player in Information and Communication Technologies. Its research activities are focused on Digital Systems with major societal and economic stakes: Embedded Systems, Ambient Intelligence and Information Processing. With its 650 researchers, engineers and technicians, the CEA-LIST performs innovative research in partnership with major industrial players in the fields of ICT, Energy, Transport, Security & Defence, Medical and Industrial Process. For more information, visit http://www-list.cea.fr.

Design Architectures & Embedded Software research activity is shared between CEA-LETI and CEA-LIST through a dedicated division. More than 240 people are focusing on RF, digital and SoC, imaging circuits, design environment and embedded software. Theses researchers perform work for both internal clients and outside customers, including Nokia, STMicroelectronics, Sofradir, MicroOLED, Cassidian, Trixell, Kalray, Delphi, Renault, Airbus, Schneider Electric, Magillem, etc...
Contents

Page 5
Thierry Colette
> Interview
*Head of Architecture & IC design, Embedded Software Division*

Page 7
Key Figures

Page 9
Scientific Activities

Page 11
Architecture & IC Design for RF & mmW

Page 21
Architecture & IC Design for Image Sensors

Page 31
Architecture, IC Design & Control for Digital SoCs

Page 49
Architecture & IC Design for Emerging Technologies

Page 57
Embedded Software

Page 69
Reliability & Test

Page 77
PhD Degrees Awarded
Interview with Thierry Collette,
Head of Architecture & IC Design, Embedded Software Division

Dear reader,

30 years ago, with the microelectronic revolution, raised the communication new age. In parallel began another one: the computing science revolution. Now, we move into a new one, which is finally the synthesis of both. With the success of the Internet, and the new needs, e.g. in health, transportation or security, more and more computing devices are being smart and connected, leading to new research fields: the efficient big data management, and the integration of hardware and software know-how inside integrated embedded systems. Furthermore, what we assist today with smartphones, tablets or onboard computers, will be widely spread to many kinds of devices: the Internet of Things is emerging.

Our multidisciplinary platform dedicated to Integrated Circuit Design and Embedded Software, allows us to address this new trend. By joining these two fields of know-how, CEA is the one of first research organization in Europe to support such an original and global offer to the industry providing a wide range of capabilities, oriented towards the applicative analysis and the exploration of integrated embedded architectures.

This platform includes the tools, methods and human competencies from front-end & back-end integrated circuit design (digital, analog & mixed) in the most advanced technologies, complex circuit emulation, hardware/software integration, to industrial test and reliability.

We hope reading this scientific report will convince you that this wide spectrum platform brings specific innovations, creating new opportunities to fulfill our first mission: support and promote the industry by innovation and technology transfer.

Thierry Collette
2012 Key Figures

3 locations:
MINATEC campus (Grenoble)
Integration Research Center (Gières)
PARIS-SACLAY Campus (Palaiseau)

160 Permanent researchers,
65 PhDs and Post-docs

Full suite of IC CAD tools,
Hardware Emulators,
& Test equipments,
for Analog, RF & Digital circuits.

37 granted patents
29 papers, journals & books
136 conferences & workshops

34MC budget
85% funding from contracts

Credits © CEA-Leti / CEA-List
Scientific Activity

Publications
165 publications in 2012, including journals and Top conferences like ISSCC, VLSI Circuits Symposium, DAC, DATE, PIERS, ESSCIRC, RTSS and ESWeek.

Prize and Awards
IEEE SOI Conference 2012 Best Paper Award granted to Olivier Thomas et al.
HIPEAC Paper Award granted to Antoine Joubert et al. for their DAC 2012 paper
ATC 2012 Best Student Paper Award granted to Ngoc-Mai Nguyen

Experts
31 CEA experts: 2 research directors, 2 international experts
9 Researchers with habilitation qualification (to independently supervise doctoral candidates)
2 IEEE Senior Members

Scientific Committees
Editorial Boards: Journal of Low Power Electronics,
19 members of Technical Programs and Steering Committees in major conferences: ISSCC, ESSCIRC, DAC, DATE, ESWEEK, RTNS, IJCNN, IWANN, EMSOFT...
Normalization committee: AUTOSAR (Automotive Open System Architecture)

Conferences and Workshops organizations

International Collaborations
Collaborations with more than 20 universities and institutes worldwide
Caltech, University of Berkeley, University of Columbia, Carnegie Mellon University, EPFL, CSEM, UCL, Polito Torino, KIT, Chalmers University, Tongji, ...
Architecture & IC Design
For RF & mmW

Wireless Sensor Node
UWB Localization
Power Amplifiers
RF & mmW Passives
RF BIST
Simulation Infrastructure for Energy Autonomous Wireless Sensor Networks with Sense & React Capability


C. Bernier, A. Didioui, D. Morche, O. Sentieys (IRISA)

ABSTRACT: Papers [1,2] present the simulation framework developed within the GRECO (GREen Communicating Objects) project. GRECO project aim is to design an energy efficient wireless platform that is totally autonomous thanks to energy harvesting capabilities and adaptive power management. To reach this goal, GRECO partners are developing a simulation framework that will allow a complete modeling of the platform in order to evaluate different power optimization strategies leading to energy neutral operations.

The huge variety of Wireless Sensor Network (WSN) applications, ranging from environmental monitoring to healthcare and smart homes, requires modular and reconfigurable platforms. Additionally, these systems have to be low cost, comply with severe size constraints, and demonstrate very high energy efficiency since the most important constraint in WSN’s remains the energy consumption. Several technologies have been developed for harvesting energy from our surroundings such as solar, wind and vibration energy. As the environmental energy can be scavenged for as long as desired, if a Power Manager (PM) is designed such that the consumed energy remains lower than the harvested energy over a long period, thus leading to an energy neutral operation (ENO), the system can theoretically reach infinite lifetime.

As the energy used by the radio transceiver represents the major part of the energy consumed by a WSN node, adaptive MAC layer policies are an active field of research. For example, in Fig. 1, the power manager, considered as the core of the energy harvesting wireless sensor node, controls the wake-up period of the microcontroller and the radio transceiver according to the harvested energy, hence keeping the node in energy neutral operation.

Clearly, the development of novel applications and deployment scenarios based on such adaptive platforms must be assisted by the simultaneous development of a dedicated simulation framework. Contrary to existing network simulators, this framework must be able to model (1) the energy harvesting subsystem which is highly dependent on time and environmental factors, (2) the cross-layer adaptive power management techniques and their associated power cost, and (3) a detailed interference model for the radiofrequency (RF) environment. Indeed, this last point is crucial in the context of a physical (PHY) layer with sense & react capability. Indeed, considerable power savings can be obtained when an RF transceiver is able to instantaneously adapt its level of performance to the time-varying conditions of the propagation channel.

Since interference (Fig. 2) can lead to packet data loss, missed alarms, delay, loss of synchronization, etc., many authors have investigated its impact on WSNs. However, none has studied the problem of interference due to intermodulation which is caused by the nonlinearity of the RF receiver. Unfortunately, linearity typically comes at the cost of increased power consumption.

We therefore propose the following new SINR model [2] for the investigation of performance degradation of WSNs under intermodulation interference:

\[
\text{SINR}_{i,j}^{\text{IMD}} = \frac{P_j}{N_i + \sum_{k \neq i, j} \alpha_{i,k} \cdot P_k + \sum_{k \neq i, j} \sum_{l \neq k, j} \beta_{i,k,l} \cdot P_k P_l^2}
\]

This model has been implemented in the GRECO simulation platform hence enabling the study of different dynamic power/performance tradeoff strategies with the aim of specifying a new sense & react transceiver for perpetually powered autonomous sensor networks.

References:
Robust and Precise Localization with Double Quadrature Receivers

Research topics : UWB, Localization, Beamforming, Antennas

F.Bautista, D. Morche, G.Masson, F.Dehmas, S.Bories

ABSTRACT: In this work, several refinements have been added to the localization techniques in impulse radio in order to improve the precision, the range as well as the robustness of the existing techniques. The receiver architecture exploited in this approach is the double quadrature. This solution has shown its capability to reach fine ranging precision in the cm range [1]. Then a multi-antenna approach has been exploited to extend the range of the receiver and to extract the Angle-of-Arrival information. More recently, we have shown that double quadrature architecture shows better robustness to antenna characteristics than the classical single quadrature [4]. Lastly, a new approach has been proposed to recover the same performances with single quadrature.

The needs for surveyed positions in civil safety and military applications require a new generation of Impulse Radio-Ultra Wide Band (IR-UWB) technology for range up to several km, capable of communication, precise localization and low consumption. Up to now, most of the IR-UWB localization solutions were based on non-coherent receivers with poor performances. In [1], as far as we know, we presented LORELEI, the first IR-UWB receiver working in the authorized 3-5 GHz frequency band and reaching a ranging accuracy lower than 10 cm. The fine localization is obtained thanks to the double quadrature architecture. High flexibility capability to cope with various channel conditions and to reduce synchronization phase has been reached thanks to the sampled baseband architecture. Even if several hundred meters range can be obtained with this solution, it may be desirable to extend the range even more as well as the localization performances.

In [2] and [3] we have exploited a multi-antenna scheme to enhance the performances. By using four antennas and LORELEI ICs, we can achieve some beamforming functionality by a simple digital algorithm. It increases the range and can be exploited to reduce the power of the unwanted blockers. It can be also exploited to extract independently the Angle-of-Arrival of each path of the impulse channel responses. The error is lower than 2 degrees over a wide angle range. This functionality opens the door of new localization algorithms which can combine Angle-of-Arrival and time of arrival for all distinguishable paths.

In [4], we have shown that the performance of the classical single quadrature receiver degrades when faced to such phenomena. On the other hand, in the LORELEI receiver, the signal is projected on an orthogonal base of two signals. As a consequence, with a 0.6cm worst case error, the system appears to be really robust against some deviation of the antenna characteristics. This emphasizes the key impact of dedicated and innovative architectures [5] to reach high performances in IR-UWB systems. Thanks to this approach, the obtained performance is among the most interesting in the state of the art [6]. More recently, we have shown that by modifying the processing done in single quadrature receivers, it is possible to reach the same robustness and precision, at the cost of an increased complexity. The next step will be to reach mm ranging precision, in order to be able to consider a wider range of applications.

References :

Figure 1 : LORELEI Architecture

Figure 2 : Received signal for 40° and 145° elevations
Design of a Fully Integrated CMOS Self-Testable RF Power Amplifier Using a Thermal Sensor

ABSTRACT: This research work presents a wideband RF power amplifier (PA) dedicated to 2GHz applications integrating a contact-less temperature sensor that allows on-chip observation and testing of the PA. Indeed, based on the static and dynamic local temperature changes caused by the PA operation, the thermal sensor can sense parameters such as output power or efficiency. This principle is applied to a 65nm CMOS PA with an OCP1 of 21dBm. We demonstrate that the output voltage of the thermal sensor follows the PA efficiency under single tone input signal conditions.

Testing issues, mainly its cost, is becoming crucially important for the success of RF SoC products for mass markets. Test cost is directly related with testing time and cost of test equipment. One strategy to enhance yield and to ease RF test consists in incorporating sensors on chip that measure the operation of the circuit-under-test (CUT). In most of the cases, the sensors imply contact to electrical nodes of the RF circuit and high-frequency signal processing, at least at the input section of the sensor. In this work we propose an alternative sensing strategy that requires no contact to the CUT since it is based on the measurement of the temperature variations in the vicinity of the circuit [1]. This technique is especially well suited for the observation of power amplifier characteristics [2,3] such as 1dB compression point or bandwidth, which can be used for testing or implementing for self-calibration loops.

The idea behind this technique is that any modification of the balance between the power drawn from the supply and the power provided to the load (or to the next stage) results in a variation of the dissipated power, that can be detected as a local temperature increases in the vicinity of the active devices of the CUT. We have applied this technique to compare some RF measurements and the results obtained with an integrated temperature sensor for figures of merit of a 2.5 GHz PA fabricated in a 0.65nm CMOS process, shown in Fig. 1. A first set of measurements is shown in Fig. 2 for a single tone input signal of a fixed frequency and varying power. When this type of signal is applied to the input of the PA, the variation of the local temperature of the PA active devices (shown in black in Fig. 2 right as measured by the sensor) tracks the variation of on-chip power, and therefore, on power delivered to the load. The plot shows how the DC value of the sensor follows closely the PA Efficiency figure of merit. A second set of measurements is shown in Fig. 3. There, the input signal consists in two tones of fixed spacing (10 kHz) and varying frequency. As the two tones are swept over the PA bandwidth the thermal signal observed at the two tones beat frequency tracks the PA bandwidth, with accuracy comparable to the conventional RF measurement. These experiments demonstrate the potentials of non-invasive, temperature based observation techniques for RF circuits BIST or self-healing.

References:
SOI CMOS RF Power Amplifier and Tunable Matching Network for Integrated RF Front-Ends

Research topics: SOI, CMOS, Power Amplifier, Tunable Matching, RF Front-End

A. Giry, G. Tant

ABSTRACT: A high integration level and tunable RF functions in SOI CMOS technology are key enablers to make smaller and more cost-effective RF Front-Ends. In this work, a two-stage SOI LDMOS linear PA and a SOI CMOS Tunable Matching Network have been designed and characterized. The obtained results represent a new step towards high efficiency integrated RF Front-Ends for future multimode multiband cellular applications.

Next generation wireless terminals and access points will have to handle an increased number of standards and frequency bands, which translates into great challenges and stringent requirements when looking at the RF front-end (RF-FE) section. Multiple Power Amplifiers (PA), RF switches and filters will be needed, which will result in an increased size and cost of the RF-FE section, especially if the multiple technologies (GaAs, SAW, IPD) currently required to achieve adequate performances cannot be circumvented. A higher integration level is the key to make smaller and more cost-effective RF-FE, and SOI CMOS technology provides an attractive trade-off among performance, cost and integration capability. The proposed research work aims at investigating SOI CMOS technology for the design of highly integrated RF-FE with reduced power consumption. To meet the needs of future cellular RF-FE, a watt-level SOI LDMOS PA with high linearity and efficiency has been developed [1] together with a low-loss SOI CMOS Tunable Matching Network (TMN) allowing improved energy efficiency under various mismatch conditions. The proposed PA and TMN have been implemented in a 0.13um SOI CMOS industrial process with a high resistivity substrate. Fig. 1 shows a micrograph of the two-stage PA which occupies an area of 0.84mm² and has been designed by using a high voltage LDMOS power device to get high efficiency and Through Silicon Vias (TSV) [2] for efficient ground connection. At 900MHz under 3.6V supply voltage, the LDMOS power stage delivers up to +33.2dBm of peak power with a maximum efficiency of 60%. When tested with a 10MHz bandwidth 16QAM uplink LTE signal, the two-stage PA provides a higher linear output power of +27dBm with less than 3% EVM. Fig. 2 shows a micrograph of the SOI CMOS TMN based on integrated high-power tunable capacitors which consist in arrays of binary weighted switched-capacitors. Each tunable capacitor exhibits 32 states and has been designed to cover the range 0.7-2.8 pF with a minimum quality factor of 40 at 2.7 GHz and a maximum power rating of +36dBm. Control logic is integrated on-chip and allows the selection of appropriate capacitance values through an integrated SPI interface. The TMN circuit occupies an area of 1.6mm² and operates under 2.5V voltage supply. As can be seen in Fig. 2, the TMN is centered on 50 Ohms and provides good impedance coverage at 1.95 GHz. When combined with a miniature dual-band antenna [3], the TMN succeeds to reduce the reflection losses down to less than 0.5 dB and allows maintaining a fairly constant radiated power even in cases of strong perturbations created by a metallic plane close to the antenna.

Figure 1: SOI LDMOS PA micrograph

Figure 2: SOI CMOS Tunable Matching Network micrograph (left) and measured Smith chart coverage at 1.95GHz (right)

References:
In this work, slow-wave coplanar transmission lines (S-CPW) and standard CPW are compared through measurements up to 65 GHz. Both S-CPW and CPW lines are fabricated on an industrial CMOS 65nm SOI with high resistivity substrate. Due to the slow-wave effect, S-CPW lines achieve a high effective permittivity that reduces the wave-length. As a result, very high quality factors are achieved that show the interest of this object for millimeter-wave (mmW) circuit design.

Millimeter-wave CMOS circuits have been intensively developed in the past decade in order to respond to a growing demand for mass-market, high throughput wireless applications. The most popular approach for the matching networks and inductive components design is the use of microstrip (MS) and CPW[1]. However, CPW and MS lines suffer, at high frequency, from high losses and low quality factor because of thin metallic layers in CMOS technologies. The concept of S-CPW respond to this problem by exploiting the slow-wave phenomenon that increases artificially the effective permittivity \( \varepsilon_{eff} \). As a result, the wavelength is decreased and thus, the corresponding physical line length for a given phase shift is reduced. This is illustrated in equation (1):

\[
Q = \frac{\beta}{2\pi} = \frac{\varepsilon_{eff} \varepsilon_0}{\varepsilon_0 c_0}
\]

Where \( \alpha \) is the attenuation constant, \( \beta \) is the spatial phase velocity and \( c_0 \) the speed of light in the vacuum.

Figure 1 shows a 3-D schematic structure of a S-CPW line integrated in a CMOS back-end with six Copper metal layers and one Aluminum top layer. It consists of a conventional CPW line with patterned metallic shield placed between the CPW and the silicon substrate.

Two characteristic impedance transmission lines (28 \( \Omega \) and 65 \( \Omega \) respectively) were fabricated in S-CPW and conventional CPW. Measurements were carried out up to 65 GHz using a two-port VNA. The extracted quality factor for each measured line is shown in figure 2.

Thanks to the enhanced effective dielectric permittivity, the quality factors of the S-CPW are significantly improved, compared to CPW ones. The Q-factors of S-CPW are increased by a factor 4 and 2.5, for 28 \( \Omega \) S-CPW and for 65 \( \Omega \) S-CPW, respectively.

In this work, high performance slow-wave lines fabricated in an advanced 65 nm HR-SOI CMOS technology were characterized and optimized. Experimental results show that the performance improvement of S-CPW, compared to conventional CPW, is mainly due to the increase of the effective permittivity. At 60 GHz, the attenuation constant of S-CPW is reduced by almost 40% and the effective relative permittivity is two to six times higher, leading to almost two to four times higher quality factor.

References:
On the Electrical Properties of Slotted Metallic Planes in CMOS Processes for RF and Millimeter-Wave Applications

Research topics: Interconnections, RF and mmW ICs

J.L. González, B. Martineau (STMicroelectronics), D. Belot (STMicroelectronics)

ABSTRACT: This research work is focused on the effects of slotted metallic planes in passive structures built using CMOS processes for RF and millimeter-wave (mmW) applications. The impact of holes on the reference plane resistance and in the capacitance of any surrounding structure to the plane are investigated through electromagnetic (EM) simulations. Two analytical expressions are derived that capture the holes impact on the plane resistivity and on the dielectric constant of the materials found between the plane and the surroundings. These expressions are used to propose a simplified EM simulation methodology for on-chip microstrip transmission lines.

Recent realizations of integrated radios operating at millimeter wave frequency (several tens of GHz) [1,2], and to a lesser extent at RF frequencies (several GHz), require the use of distributed passives such as transmission lines. These structures must be fabricated by respecting strict manufacturing rules imposed by the semiconductor processing tools and procedures. For large area metallic surfaces that are required to build the reference ground planes of microstrip lines, for example, the manufacturing rules impose a maximum density of metal, so that such planes must be pierced with holes, as indicated in Figure 1.a. Up to now, little attention was paid on the impact of this modified planes with respect to the ideal, continuous metallic plane that should be used if possible. Figure 2.b shows the basic parameters of a section of a slotted plane. A basic cell consisting on a section of the plane with a single hole can be defined, and the plane can be considered as a 2D repetition of this basic cell. The relative size of the hole with respect to the size of the basic cell sets the basic parameter for the plane: the metal density (or its inverse, the hole density).

In this research work we have analyzed the modification of the electrical properties of the plane that are caused by the presence of the hole, in comparison to an ideal, continuous plane without holes (i.e., with a 100% metal density). Figure 2 shows how the conductivity of the plane is reduced by a factor of 5 if the metal density is reduced up to a 40%. This modification of the plane conductivity is observed for different plane thicknesses, such as those obtained by using the various metallization levels available in CMOS processes. The holes opened in the plane also modify the electric fields of the surrounding structures to the plane, such as for example the capacitance of a line to the plane, where relative changes by a factor of 3 are observed.

The observed significant change in the plane and transmission line properties observed must be taken into account for an accurate design of this type of structures. In [3] a simulation strategy is proposed that goes in that direction.

References:
BAW Filters for Ultra-Low Power Narrow-Band Applications

Research topics: BAW, Ultra-Low-Power, Narrow-band

C. Bernier, J.-B. David

ABSTRACT: This paper presents an original method for the design of Bulk Acoustic Wave (BAW) filters for a new class of applications: ultra-low-power, narrow-band RF filtering. To this end, a filter co-design methodology, based on existing BAW resonator technology and fabrication processes, is developed and linked between decreasing filter bandwidth and decreasing power consumption of the associated integrated circuit is demonstrated. Depending on required bandwidth, the power dissipation of the driving electronics can be reduced by large factors (10 to 40). High-IF receivers and Wake-up receivers are main applications.

In addition to classical out-of-band spurious signals, with the increasing number of RF standards and devices, susceptibility to in-band blockers is becoming a serious issue in wireless systems. This is especially true in the context of shared RF bands such as the 2.4GHz ISM band. Faced with strong in-band blockers, the RF system architect has the choice to either increase the linearity specification of the receiver, which requires power consumption, or to increase the receiver selectivity, and this ideally as close as possible to the antenna. There is therefore a need for narrow-band filters at VHF/UHF frequencies [2]. Whereas achieving this with active designs is prohibitive in terms of power consumption, BAW filters [3] appear to be good candidates if they can be made sufficiently narrow-band and if the problem of frequency agility is solved e.g. using a high intermediate frequency (IF) architecture.

Most IC blocks in low power architectures are designed to optimize the voltage transfer characteristic and, for this reason, high input and output impedances are favored. However common RF filters impedance (50Ω) is a difficulty in low power designs, meaning current dissipation. The aim of this work is therefore to simultaneously solve these two problems of narrow required bandwidths and low power dissipation. To do so, we used a co-design approach to simultaneously optimize the filter response while minimizing the power consumption of the IC design, leading to an original high impedance BAW filter topology.

Instead of classical S parameters, the approach is based on the notion of "equivalent load" which is convenient for both the BAW filter and IC designers. Considering the equivalent circuit shown in fig.1, we are then allowed to write the following expression for voltage gain $G_v = \frac{V_{out}}{V_{in}}$:

$G_v = \frac{-G_m \cdot Z_{21} \cdot Z_{out}}{Z(Z_{out} + Z_{in} + Z_{22} - Z_{11} - Z_{21} - Z_{12})}$

In the particular case where $G_m = 1$, this expression is homogenous to an impedance, an "equivalent load", which is directly related to the $S_{21}$ scattering parameter of the BAW. We converged to a filter response which has both a narrow bandwidth and a large equivalent load, by drastically reducing the frequency spacing between the series and parallel resonators, to the overlap area (see fig.2), which is forbidden in classical power transmit BAW filter design.

Note that these filter responses are obtained with resonator parameters extracted from state of the art devices. The only difference with respect to existing process flows is the loading layer thickness which must be modified to create smaller frequency offsets. The filter design methodology described in this work has been explored in High-IF architecture for ISM band. It can also be used to design an extremely selective, ultra-low power RF gain stage for a wake-up radio where typical input and output impedances are small capacitances (e.g. 100fF), allowing the design of a lattice filter with equivalent load greater than 1kΩ and BW-3dB<4MHz.

References:
A Frequency Measurement BIST Implementation Targeting GigaHertz Applications

Research topics: Design for test, Radiofrequency measurements

M. Dubois, E. de Foucauld, C. Mounet, S. Dia (Presto Engineering) and C. Mayor (Presto Engineering)

ABSTRACT: We propose a Built-In Self-Test (BIST) technique for measuring the natural resonance frequency of oscillators which are set much higher than the working speed of current Automated Test Equipment (ATE). Based on an asynchronous counter, the BIST response corresponds to a digital output code proportional to the frequency of the oscillator under test. The efficiency of the proposed BIST is demonstrated on an Ultra-Wideband transceiver, whose communication frequency ranges in the band of 7.25GHz to 8.5GHz.

We present a BIST dedicated to the measurement of a high frequency transceiver based on the Super Regenerative Oscillator (SRO) principle. The suggested BIST architecture relies on an asynchronous counter that deduces the frequency measurement counting the number of oscillation periods within a given period of time. The digital output test response suits any digital communication and processing systems, which can use the information for test purposes and/or self-calibration or compensation techniques.

For respecting standardization rules, this frequency is set to F=7.875GHz, the center of the 7.25GHz to 8.5GHz UWB band. For this application, the optimal BER is reached when the oscillator resonance frequency matches the central frequency of the input signal.

The super-regenerative receiver with its BIST is implemented in a 0.13µm CMOS technology. The connection of the output of the SRO and the input of the BIST has to be as short as possible to limit the parasitic capacitance of the net connection and crosstalk with other signals. On the other hand, the introduction of the BIST close to this critical path of the system increases the risk of performance reduction of the system even when the BIST is switched off.

A close ground line could increase crosstalk in the reception path. Therefore, only the first stage is introduced very close to the SRO whereas the next stages are located as far as possible of the output of the SRO.

Figure 1 represents the SRO resonance frequency obtained by the measurement with the ATE resources FMeas and with the BIST technique FBIST. This scatter plot shows the excellent correlation between both measurements.

Figure 2 shows the chip connected in its socket. The SRO output is connected to the mixer through the coaxial cable whereas the BIST used only digital pins of the tester.

Figure 2: Picture of the chip connected to the ATE Verigy 93K.

Figure 1: Frequency measurement of the SRO resonance with the ATE resources and the BIST technique.

Reference:
Architecture & IC Design
For Image Sensors

High Performance IR Imagers
3D Integration for Imagers
Advanced Integrated Algorithms
An 88dB SNR, 30µm Pixel Pitch Infra-Red image Sensor With a 2-step 16 bit A/D Conversion

Research topics : CMOS image sensors, Infra-Red, pixel-level ADC


ABSTRACT: A new readout IC (ROIC) with a 2 step A/D conversion for cooled infrared image sensors is presented in this paper. The sensor operates at a 50Hz frame rate in an Integrate-While-Read snapshot mode. The 16-bit ADC resolution preserves the excellent detector SNR at full well (~3GeV). The ROIC, featuring a 320x256 array with 30µm pixel pitch, has been designed in a standard 0.18µm CMOS technology. The IC has been hybridized (indium bump bonding) to a LWIR (Long Wave Infra Red) detector fabricated using our in-house HgCdTe process. The first measurement results of the detector assembly validate both the 2-step ADC concept and its circuit implementation. This work sets a new state-of-the-art SNR of 88dB.

Used in security and defense applications, cooled (77K) Infrared HgCdTe (Mercury Cadmium Telluride) hybrid sensors (detector bump-bounded over the CMOS IC) are very demanding in terms of SNR (typical state-of-the-art values are in the 70-80dB range). The detector sensitivity can be limited either by the incident number of photons during one frame or by the CMOS readout IC (ROIC) charge handling capacity. In many thermal imaging conditions, this second point predominates. This limited charge well capacity is determined by two CMOS process constraints: the integration capacitance that has to be fit in a given pixel area and voltage range. To overcome this limitation, new ROIC architectures must be developed. Pixel-level analog-to-digital conversion is a very attractive solution that enables high dynamic range imaging and SNR breakthrough performance, while being compatible with an IR pixel size.

The overall architecture of the sensor is given Fig. 1. At the end of the integration time, the global shutter pixel delivers an 11 bit digital output as well as its residue analog output. This residue is then converted using a 5 bit flash ADC, which gives a final 16-bit digital output. In order to make the Integrate While Read (IWR) feature possible, a whole image memory (SRAM) is needed.

As illustrated on Fig. 3, the pixel uses a pixel-level ADC technique that is described in [1]. It consists in counting charge packets so that, at the end of the integration time, the pixel counter contains a digital value proportional to the total integrated charge and the residue remains on the integration capacitance (Cint). GS (Global Shutter) is a global signal while RS (Row Select) is a line-wise signal that allows the pixel to write on the digital bus on one hand and on the analog bus on the other. For a fixed resolution of 16 bits, the number of bits at the pixel level can be assessed on an area criterion. For the 0.18µm process we used, Fig. 2 shows that there is a tradeoff between the integration capacitance and the counter depth.

The test chip was fabricated in a 1P6M 0.18µm standard CMOS process. This 320x256 hybrid HgCdTe sensor demonstrates how the 3Ge- full well capacity associated with a 16-bit ADC resolution paves the way for a breakthrough in thermal sensitivity. In electro-optical tests, a peak SNR of 88dB has been reached with power consumption below 72mW.

| Research topics: CMOS image sensors, Infra-Red, pixel-level ADC
| ABSTRACT: A new readout IC (ROIC) with a 2 step A/D conversion for cooled infrared image sensors is presented in this paper. The sensor operates at a 50Hz frame rate in an Integrate-While-Read snapshot mode. The 16-bit ADC resolution preserves the excellent detector SNR at full well (~3GeV). The ROIC, featuring a 320x256 array with 30µm pixel pitch, has been designed in a standard 0.18µm CMOS technology. The IC has been hybridized (indium bump bonding) to a LWIR (Long Wave Infra Red) detector fabricated using our in-house HgCdTe process. The first measurement results of the detector assembly validate both the 2-step ADC concept and its circuit implementation. This work sets a new state-of-the-art SNR of 88dB.
| Used in security and defense applications, cooled (77K) Infrared HgCdTe (Mercury Cadmium Telluride) hybrid sensors (detector bump-bounded over the CMOS IC) are very demanding in terms of SNR (typical state-of-the-art values are in the 70-80dB range). The detector sensitivity can be limited either by the incident number of photons during one frame or by the CMOS readout IC (ROIC) charge handling capacity. In many thermal imaging conditions, this second point predominates. This limited charge well capacity is determined by two CMOS process constraints: the integration capacitance that has to be fit in a given pixel area and voltage range. To overcome this limitation, new ROIC architectures must be developed. Pixel-level analog-to-digital conversion is a very attractive solution that enables high dynamic range imaging and SNR breakthrough performance, while being compatible with an IR pixel size.
| The overall architecture of the sensor is given Fig. 1. At the end of the integration time, the global shutter pixel delivers an 11 bit digital output as well as its residue analog output. This residue is then converted using a 5 bit flash ADC, which gives a final 16-bit digital output. In order to make the Integrate While Read (IWR) feature possible, a whole image memory (SRAM) is needed.
| As illustrated on Fig. 3, the pixel uses a pixel-level ADC technique that is described in [1]. It consists in counting charge packets so that, at the end of the integration time, the pixel counter contains a digital value proportional to the total integrated charge and the residue remains on the integration capacitance (Cint). GS (Global Shutter) is a global signal while RS (Row Select) is a line-wise signal that allows the pixel to write on the digital bus on one hand and on the analog bus on the other. For a fixed resolution of 16 bits, the number of bits at the pixel level can be assessed on an area criterion. For the 0.18µm process we used, Fig. 2 shows that there is a tradeoff between the integration capacitance and the counter depth.
| The test chip was fabricated in a 1P6M 0.18µm standard CMOS process. This 320x256 hybrid HgCdTe sensor demonstrates how the 3Ge- full well capacity associated with a 16-bit ADC resolution paves the way for a breakthrough in thermal sensitivity. In electro-optical tests, a peak SNR of 88dB has been reached with power consumption below 72mW.

Table 1: Summary of the sensor features vs other works

<table>
<thead>
<tr>
<th>Power/pixel (mW)</th>
<th>Peak SNR</th>
<th>Pixel pitch</th>
<th>CMOS process</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5µW</td>
<td>88dB</td>
<td>30µm</td>
<td>0.18µm</td>
<td>320x256</td>
</tr>
<tr>
<td>1.7µW</td>
<td>85dB</td>
<td>50µm</td>
<td>0.35µm</td>
<td>128x128</td>
</tr>
<tr>
<td>10µW</td>
<td>75dB</td>
<td>30µm</td>
<td>0.18µm</td>
<td>16x1</td>
</tr>
<tr>
<td>9.7µW</td>
<td>70dB</td>
<td>50µm</td>
<td>0.18µm</td>
<td>64x64</td>
</tr>
</tbody>
</table>

References:
Linear Photon-Counting with HgCdTe APDs

Research topics: photon counting, image sensor, infrared

F. Guellec, G. Vojetta, J. Rothman

ABSTRACT: A custom readout IC has been developed for photon counting. It features a pixel with 115μV/e- conversion gain, 10e- noise and 13μW power consumption to comply with upcoming integration in focal plane arrays. It targets single-photon detection at moderate APD gain. This circuit will enable operation in the Short Wave infrared (SWIR) band where the APDs exhibit a significantly reduced gain for a given photodiode bias voltage compared to the Mid Wave Infrared (MWIR) band. The use of SWIR APD allows reducing the Dark Count Rate (DCR) or increasing the operating temperature (for a given DCR).

In the pixel, the photodiode current pulse is integrated on the input node capacitance. It is interesting to use this relatively small capacitance (around 15fF) to perform a fast and low-noise current to voltage conversion. The resulting voltage is then amplified by 10 with a low-noise stage having a small input capacitance compared to the photodiode junction capacitance in order to keep the total input node capacitance as low as possible. The pixel power consumption was reduced to 13μW in order to allow its use in focal plane arrays.

Single and proportional photon detection capability was characterized by measuring the amplitude distribution of the circuit output voltage step occurring with a laser impulse.

Infrared avalanche photodiodes (APD) using HgCdTe compound semiconductor material are developed by CEA-Leti since several years. These photodiodes are typically cooled at 80K and operate below the breakdown with an avalanche mechanism only initiated by electrons resulting in a linear amplification (M) with a very low excess noise factor (F).

For this work [1], we developed a new low-noise readout electronic circuit in a standard 0.18μm CMOS technology. It targets single-photon detection at moderate APD gain. This circuit will enable operation in the Short Wave infrared (SWIR) band where the APDs exhibit a significantly reduced gain for a given photodiode bias voltage compared to the Mid Wave Infrared (MWIR) band. The use of SWIR APD allows reducing the Dark Count Rate (DCR) or increasing the operating temperature (for a given DCR).

The high conversion gain (115μV/e-), low noise (10e-) custom IC that has been developed is useful to characterize APD gain and excess noise. It allowed demonstrating linear mode photon counting at low APD gain (40) with an estimated 90% PDE and 800kHz DCR for a threshold at 40% of average single-photon amplitude for MWIR APD at 80K.

References:
A low-noise, 15µm pixel-pitch, 640x512 hybrid InGaAs image sensor for night-vision

Research topics: image sensor, infrared, night vision

F. Guellec

ABSTRACT: This paper presents the design of a 15µm pixel-pitch, 640x512 CMOS readout IC. A careful noise analysis of the C-TIA pixel circuit is necessary to achieve low noise performance with a high conversion gain. A 30e- read noise for a 71dB Dynamic Range (DR) has been reached with the developed hybrid InGaAs image sensor operated in rolling shutter with Correlated Double Sampling. Experimental results are in good agreement with simulated values. In high gain configuration (17.6µV/e-) a read noise of 30e- has been reached for a dynamic range of 71dB. In low gain configuration (1.9µV/e-) we get a DR of 79dB. As expected, the lower noise floor in high gain is obtained at the expense of the dynamic range. This trade-off should be adjusted according to application needs. The dual gain functionality of the pixel furthermore enables both night and day use. The 640x512 image sensor operates at a frame rate up to 120fps with a total power consumption of 150mW.

Hybrid InGaAs infrared detectors allow easy and compact camera integration as cooling is not needed. They are sensitive from the Short Wave infrared (SWIR) (λ=1.7µm) down to the visible (λ=0.4µm) when the substrate is thinned. The SWIR band presents some key advantages for night vision. In this band, the haze offers a good transmission and an optical phenomenon occurring in the atmosphere (called airglow or nightglow) causes a weak generation of light.

In this context, we developed in collaboration with the III-V Lab a low-noise, 15µm pixel-pitch, 640x512 hybrid InGaAs image sensor for night vision [1, 2]. We were in charge of the readout IC design in a standard 0.18µm CMOS technology. The pixel is based on a dual gain C-TIA circuit with an anti-blooming function. The image sensor is operated in rolling shutter with an optional correlated double-sampling mode which is useful to reduce the noise in high-gain configuration. Thanks to a thorough noise analysis (taking into account power supply noise and CDS filtering) and careful circuit optimization with respects to area and power consumption constraints state of the art performances have been reached.

Experimental results are in good agreement with simulated values. In high gain configuration (17.6µV/e-) a read noise of 30e- has been reached for a dynamic range of 71dB. In low gain configuration (1.9µV/e-) we get respectively 108e- and 79dB. As expected, the lower noise floor in high gain is obtained at the expense of the dynamic range. This trade-off should be adjusted according to application needs. The dual gain of the pixel allows a use in both night and day conditions as well as image fusion if needed. The 640x512 image sensor operates at a frame rate up to 120fps with a total power consumption of 150mW.

Further work is carried out to reduce the pixel pitch to 10µm while maintaining good noise performance in the aim of developing a future 1280x1024 detector.

Figure 1: Simplified pixel architecture and modeled noise spectral density (dashed blue: input noise, bold blue: output noise, red: output noise after CDS)

Figure 2: View of the packaged hybrid image sensor and picture taken with the developed camera.

Références:
ABSTRACT: This paper presents a High Dynamic Range (HDR) image sensor architecture that uses capabilities of three-dimensional integrated circuit (3D IC) to reach a dynamic range over 120 dB without modifying the classic (3T or 4T) pixel architecture. The integration time is evaluated on subsets of pixels on the lower IC of the stack and then sent back by vertical interconnections to the sensor array. This work evaluates the performance of an analog Winner Take All circuit, used to detect the maximum exponent corresponding to the optimum integration time chosen for every group of pixels.

Integrating more complex functions within the same circuit is one of the main quests for the microelectronics industry. Three-dimensional integration by circuit stacking (3D stacking) constitutes a promising way to achieve this goal. It allows notably pushing some limitations that circuits have reached nowadays. The main motivation is to take advantage of the 3D topology to exceed the limited dynamic range of the standard image sensors while keeping the classic 3T or 4T pixel architecture. This work presents a new architecture of an image sensor that allows reaching a dynamic range over 120dB without modifying the classic (3T or 4T) pixel architecture. This architecture takes advantage of emergence of technologies of dense vertical interconnections, Through Silicon Via (TSV), to locally adapt the integration time of a group of pixels. The coding of a high dynamic range and a high PSNR image leads to an increase of the data throughput, at the IOs of the circuit. The HDR architecture is so coupled to a two-level compression system [1, 2].

To mitigate the available lack of pixel area and TSV pitches of about tens of microns, the circuit proposed in this work takes advantage of 3D stacking of 2 integrated circuits. The circuit consists of two stacked dies vertically interconnected by TSVs. The upper die performs image acquisition and is based on the architecture of a classical 2D image sensor, with 3T or 4T pixels. The processing performed on the lower die contains two stages. Firstly, it estimates the best suited integration time for every macro-pixel, and then, generate the command signal that adjusts the integration time. To deduce the optimal integration time, we use the circuit architecture presented in Fig.1.

In every macro-pixel, the maximal voltage drop $\Delta V$, corresponding to the minimum integration time, is determined by means of a Winner Take All (WTA). We have designed a WTA circuit in 32nm double oxide CMOS technology. Due to its analog nature, the transfer function of the WTA has an offset and a gain error. The output voltage as a function of different sets of input voltages has been simulated (Fig. 2).

The characteristic equation of the resulting curve shows a gain about 0.973 and a 321 mV offset voltage. These values are coherent with the analytical expressions of the offset (Eq.1) and the gain (Eq.2):

$$G = \frac{V}{V} \left(\frac{S_m 1}{m + \sum_{d=1}^{D} \frac{S_m 1}{d + S_d 2^d}}\right)$$

$$V = \left(\frac{\it{g}}{\it{g} - \it{V}_W V - 1} \ nV\right)$$

References:
Computational SAR ADC for a 3D CMOS Image Sensor

Research topics: 3D integration, CMOS image sensor, image descriptor

A. Verdant, A. Dupret, M. Tchagaspanian, A. Peizerat

ABSTRACT: The architecture and simulation of a Computational SAR ADC (C-SAR) dedicated to the processing of image descriptors for a 3D CMOS image sensor are reported here. The differential charge sharing architecture enables to A/D convert the convolution of multiple binary weighted pixel signals on multi-scale kernels. The CMOS image sensor is constituted of two tiers (two 3D layers). An array of C-SAR is implemented on the bottom layer. Each C-SAR is associated to a square of 8×8 pixels on the top layer, with a pitch of 10µm and a fill factor of 80%. The total noise of 460µVRMS simulated at transistor level on a 65nm technology enables to reach a processing resolution of 9 signed bits on 0.5V pixels dynamic, with a FOM of 6.25pJ/pixel.

In automotive applications, the driver drowsiness detection is extremely constrained in terms of processing bandwidth. The eye blinking analysis is based on high frame rate video (200fps). The general principle of the method used to extract blink features from video. A part of this processing relies on the face detection from Viola-Jones algorithms using Haar-like descriptors. Despite of the high-throughput architectures associated to standard CMOS image sensors allowing spatial weighted sums (convolution) to be computed, integrated processing features are mandatory to reduce power and silicon area costs. Hence, to overcome the limitations associated to the use of DSP, processing features have been successfully implemented in CMOS image sensors [1].

The Computational-SAR (C-SAR) architecture allowing the calculation of the Haar descriptors is here presented. This topology takes benefit from the high bandwidth of the SAR ADC together with low power consumption. Indeed, the successive approximation converters are known to provide the best FOM considering the energy per step. This C-SAR processing unit will be exposed considering its implementation in a two tiers 3D CMOS image sensor, chosen to preserve the fill factor of the sensor array.

The C-SAR is conceived as the building block of the readout circuit of a 3D CMOS image sensor. The top tier of (Fig. 1) embeds a 32×32 macropixel array. Each macropixel is composed of a square of 8×8 10µm back-illuminated pixels being locally read in rolling shutter. On the second tier, an array of 32×32 C-SAR cells is implemented to compute the binary weighted sum of pixels. Each C-SAR cell is associated to a macropixel and is thus shared by its block of 8×8 pixels. The connection between tier 1 and tier 2 is realized in direct metal bonding, in a face to face configuration. Only one interconnection is required for each 8 pixels column of a macropixel. The readout pipeline of a macropixel is presented in Fig. 2. Each column of the macropixel in tier 1 is associated to a sample and hold circuit in tier 2. A bank of 2×8 capacitors thus enables to store the reference (black level) and pixel signal of a line of 8 pixels being read in rolling shutter mode. The sampled data are then multiplexed towards the analog to digital processing unit to be then computed.

A low power consumption architecture has been simulated for a processing resolution of 9 signed bits reaching a FOM of 6.25pJ/pixel. Compared to standard processing architectures, no additive time is required, the processing being performed together with conversion. This C-SAR is suitable with high frame rate up to 2200fps.

References:
Design and Optimization of Two Motion Detection Circuits for Video Monitoring

A. Dupret, M. Zhang, N. Llaser and H. Mathias (IEF)

ABSTRACT: In a classical video monitoring system, though for most of the time the captured images contain no relevant information, it cannot prevent the monitoring system from useless power consuming for image processing. One technique proved to be effective for a video monitoring system to reduce its power consumption is based on macro-pixels or blocks of pixels for region of interest (ROI) finding. The key feature necessary for ROI finding corresponds to motion detection. To be able to implement ROI detection, two CMOS based motion detection circuits are proposed. Both are designed and optimized to have fewer transistors, lower power consumption, higher sensitivity and better uniformity detection.

Security issue has been becoming more and more important in our society. Video monitoring system offers one of the solutions. However, statistically speaking, most of the images captured by a video monitoring system contain no relevant information. The data transmission, the image processing as well as the power consumption must be performed, which uselessly increases not only image processing time but also power consumption. To improve the video monitoring system performance, the macropixel approach proposed by our research team offers several advantages and therefore seems quite promising.

Two original motion detection circuits are designed to be integrated within the macro-pixel, which is a promising technique we have proposed to achieve a high resolution / a low power consumption video monitoring system. The design is made in a CMOS 0.35μm technology. Simulation results show that both circuits display low power consumption and dispose a variable threshold choice for motion detection by varying electrically the comparator window width. However, the second circuit clearly exhibits more interesting aspects than the first one in terms of fewer transistor numbers, lower power consumption and much lower non-homogeneity within the designed common mode input range as well as much higher voltage gain. The next step consists in integrating both proposed circuits on silicon to experimentally evaluate their functionality.

References:
Towards a Real Time Sensor for Focusing Through Scattering Media

Research topics : Image sensor, Wavefront correction

ABSTRACT: Materials such as milk, paper, white paint and biological tissue scatter light. As a result, transmitted light intensity through these materials is a speckle pattern, having often a short persistence time. Recently, advances in optics to control light through disordered media have reported an increasing efficiency. Consequently, that allows us to foresee a real time sensor that achieve such task in an integrated way. Thereby, in this perspective, we propose a genetic algorithm implemented with pyramidal approach in a CMOS image sensor, which matches integrated data processing and short persistence time. Our algorithm have been simulated with a faithful model. Results show at least a gain of a factor 10 compared to the state of the art.

Materials such as milk, paper, white paint and biological tissue are opaque due to multiple scattering of light. Consequently, the interaction between the media and the light beam causes phase changes of light. Recently, many works have been reported to control coherent light through scattering media. The principle consists in correcting phase perturbations produced by the media achieving inverse diffusion. Indeed, the use of phase only Spatial Light Modulators (SLM) for wavefront correction is a promising way to achieve focusing coherent light. Wavefront correction can be achieved by finding the optimal set of phases thanks to SLM which phases can be adjusted. This task constitutes an optimization problem.

Turbid media, especially biological tissue, often feature short persistence time, of few milliseconds. Hence, the corrected wavefront must be computed within the persistence time. This complicates the optimization process, which hence must be robust with regards to high noise level. Some works propose focusing sequential algorithms or the measurement of transmission matrix that allows generating the correct phase set that, in turn, will allow focusing the light beam.

All these algorithms are time consuming or suffer from lack of robustness in noisy environment. Recently an efficient genetic algorithm has been presented.

The parallel processing at the pixel level allows dramatic acceleration of processing [1]. A major challenge is to make the implementation compatible with pixel level. In that scope, we present a pyramidal genetic algorithm (GA) that can be implemented within a CMOS image sensor.

For instance, considering a 256x256 pixels image array, a persistence time of 2 ms, and assuming that the algorithm needs 250 frames to converge, the image sensor have to capture 125 000 frames per second (fps), corresponding to a transfer rate of nearly 9 Giga-pixels per second. The standard approach, i.e. camera and processor suffers from limitations: delay due to frame transfer and centralized data processing. Therefore, we aim at developing a dedicated smart image sensor allowing enhancing the focusing convergence time with regards to persistence time in biological media. Indeed, the standard optical setup corresponding to this model, used for testing the algorithms and simulating their implementation, is shown in Fig. 1. A Laser source illuminates a reflective SLM array. Each element of the SLM array shifts the phase of its incident light from 0 to 2Π. Next the light beam is scattered by the media, and finally the transmitted intensity is recorded on an image sensor.

In order to compare our implementation to state of the art, we consider the previously used criterion of enhancement defined as the transmitted intensity in the chosen target (focus point) over the averaged transmitted intensity before optimization. This criterion is measured with regards to the number of frames acquired by the image sensor. An example of transmitted light is shown in Fig. 2 running our genetic algorithm.

Results show at least a gain of a factor 10 with our algorithm compared to state of the art. Moreover, the pyramidal approach compared to the classical one allows at least a gain of a factor 2. Finally, our genetic algorithm has been evaluated with different noise levels and compared to the state of the art. Results show a convergence of our algorithm with high noise level while the state of art does not converge.

References :
ABSTRACT: Image sensors stabilization is usually based on accelerometers. To reduce the number of external components of digital image sensors, an integrated image stabilization system is envisaged. Such a system requires modeling the blur due to hand tremor and a general sharpness metric to quantify the gain of such a stabilization system. We aim at providing an accurate model of the hand tremor and its impact as a Point Spread Function. In order to define the specification of the image based image stabilization we have derived perceptual visual quality sharpness metric for camera shake blur. This sharpness metric is based on visual blur test. It proves to fit well ground truths such as mean opinion score database and quality ruler measure of blur.

The digital imaging market is characterized by conflicting demands: smaller pixels, in order to attain large format and to reduce the cost of the die, and sensor high performances in terms of sensitivity and signal-to-noise ratio (SNR). To keep a reasonably high SNR, longer integration times are required. Yet, longer integration time makes the quality of the resulting image sensitive to motion blur. Since hand tremor is more important for lighter device these problems are even more dramatic for compact cameras and cameraphones. Therefore, an image stabilization (IS) mechanism is to be used to reduce blur due to the camera shake. In order to get rid of the classical mechanical accelerometers used in IS, our approach is to develop an integrated image-based motion detection. The specifications of this integrated image based motion detection derive from the impact of hand tremor blur on the quality of the image. Our work so leads to a faithful model of hand tremor and a metric to measure the impact of blur on the quality of images.

The angular variations between camera and scene caused by hand tremor present a power spectral density (PSD). The characteristics of the camera (focal length, pixel pitch, etc.) are responsible for the conversion of angular tremor to the translation motion of pixels on the image sensor. The Point Spread Function (PSF) then results from the integration of the motion signal. The PSF is used to generate the motion blurred image from a reference scene by convolving it with the reference image. The particular blur induced by the hand tremor in the resultant image has not been well characterized regarding its impact on human perception. Yet, two particular types of blur (Gaussian blur and straight-line motion blur) have been studied. The Gaussian blur can be found in defocus condition while the straight-line blur is generally used as a simplified model of the motion tremor. For the Gaussian blur, some publicly available databases of subjective quality data already exist. The data base uses several distortions such as Gaussian blur providing the Mean Opinion Score (MOS). Regarding the straight-line blur quality ruler based on the just noticeable difference (JND) have been studied. Yet, due to the complexity of the camera shake, these particular results are not suitable for complex blurs.

Thus we developed a sharpness quality metric based on the PSF of the camera shake. The result of the metric is then normalized regarding both, JND and MOS database to provide direct human perceptual value not limited to the general case as straight-line blur and Gaussian blur. Our metric is also based on the circle of confusion which can take apart on the final user viewing condition (such as web applications, display, printing...). The metric is validated by user test such as image comparison and it fits the experimental trends of other databases both in the case of linear motion blur (Fig.1) and arbitrary motion blur (Fig. 2).

To our best knowledge this is the first metric that can measure all types of arbitrary blur. This metric leads to specifying image based electronic image stabilization systems and can quantify the subjective final gain of the overall IS.

References:
Saliency-Based Data Compression for Image Sensors

Research topics: visual attention compression, architecture-algorithm co-design

Tien Ho-Phuoc, L. Alacoque, A. Dupret, A. Guérin-Dugué, (GIPSA-LAB)

ABSTRACT: As saliency models have revealed ability to predict where observers fixate during scene exploration. Embedding a saliency model into an image sensor for data compression allows allocating bit-rate budget according to the saliency level of a region. This paper presents an original implementation of a saliency-based data compression algorithm and architecture. A video-rate compliant, compact saliency models is designed to allow its integration within an image sensor. It shows better performances in predicting human fixation than the state-of-the-art models. A simpler version of our proposed model requires 256 times less memory. Second, a Haar wavelet based compression is applied according to the saliency of regions in each frame.

Lossy compression algorithms enable higher compression ratio than their lossless counterparts at the expense of artifacts that are visually disturbing, especially on salient regions, and when high compression ratio are used. An image sensor integrating a saliency model is able to adapt the compression ratio according to saliency. Its implementation with the image sensor must be compliant with strong hardware constraints, i.e. limited memory and processing elements within the image sensors. We first propose a very compact - yet efficient - video saliency model that complies with the low-complexity requirement of image sensors. The proposed model combines - through the "OR" operation - motion saliency with the central fixation bias, a human viewing tendency. Motion saliency is computed in blocks thanks to an adaptive threshold (Fig. 1) resulting in little required memory (Fig. 2). The central fixation bias is constant for all frames and is stored within a look-up table. Second, the compression step is applied to each block. If a block is salient, all its information is conserved. By contrast, non-salient blocks are reconstructed by only their LL (approximation) component from the Haar wavelet transform. Only compact operators are used in the proposed model.

Fig. 3 illustrates the saliency map of the proposed saliency model - exploiting motion and the central fixation bias - for a given frame. It is also compared with the saliency maps of two other algorithms: the first one is Sigma-Delta algorithm, since it features a very compact physical implementation, the second is Itti's model that usually serves as a reference. The best performances are obtained with our model.

The proposed framework presents an original, compact yet efficient, saliency-based data compression model for image sensors. It is flexible and so might be improved by adding filtering operators.

References:
A New Approach of Smart Vision Sensors

Research topics: smart imagers, adaptive processing, feedback

J. Bezine, M. Thévenin, R. Schmit, M. Duranton, M. Paindavoine (LEAD)

ABSTRACT: Today’s digital image sensors are used as passive photon integrators and image processing is essentially performed by digital processors separated from the image sensing parts. This approach imposes to the processing part to deal with definitive pictures with possibly unadjusted capture parameters. This work presents a self-adaptable preprocessing architecture concept with fast feedback controls on the sensing level. These feedbacks are controlled by digital processing in order to adapt the exposition and processing parameters to the captured scene parameters. This innovative way of designing smart vision sensors, integrating fast feedback control enables new approaches for machine vision architectures and their applications.

Nowadays, in most image processing systems, the sensor is separated from the image processing part, pixel values being sent serially. First, photons are integrated for a predefined exposition time; next, a control circuit reads and sequentially converts the pixel values from analog to digital. Finally, pixel values are sent to an image processor for image enhancement or computer vision applications. Thus, image processing systems consider pixel values after the end of full exposure. In that way, corrections such as dynamic range enhancement or image stabilization need to be added in order to suppress the effects of unadjusted image capture parameters. This is particularly true in vision applications such as obstacle detection, or target tracking, the image sensor being used on moving vehicles, suffering from their vibrations and often analyzing difficult scenes (highly contrasted or bad weather conditions).

During the last decade, image processing systems tend to link sensing parts to the processing units. Near-pixel processing were introduced in smart sensor, at analog or digital level, in order to refine or adapt captured images before final processing, thus optimizing it. To further improve silicon and energy efficiency, this work proposes to associate even more closely image capture and image processing by adding fast and local feedback controls in the usual image capture process.

![Figure 1: Schematic of feedback integration approach in the image processing flow.](image)

This adaptation of the usual image capture process is presented in Fig. 1. It is firstly based on the close control of the image capture parameters (exposure time, conversion gain and pixel reset), during photons integration time. This introduces the use of frame sub-exposures to construct a full frame. These sub-exposures may be considered as sampled continuous readout. To deal with the control needed for our approach, we propose a hardware architecture adaptation relying on 3D stacking technologies to process pixel quickly enough to enable capture control – by feedback – during the image construction. It associates a 2D preprocessing elements matrix to the photo-sensitive layer, separated in pixel blocks. These preprocessing elements are designed to do generic vision pre-computing in order to provide a preprocessed image, or specific image features to the associated high level processing unit. The innovative purpose of this layer is to locally control the photo-sensitive layer by processing incoming pixel values on the fly, and sending back adapted capture parameters.

![Figure 2: Multi-exposure adaptation and resulting motion related Region-of-Interest delivered by the sensor integration fast feedback adaptation.](image)

This work was presented in [1], showing the first results of feedback controlled design. Fig. 2 shows an application of our approach for motion detection in a highly contrasted environment. As image processing algorithms are designed for traditional architecture that processes images after their acquisition, new algorithms must be considered in order to benefit from this smart sensor architecture. Further work will investigate such designs, and enhance our smart sensor adaptation capabilities and flexibility.

3D Architectures & Circuits
Manycores
FDSOI Circuits & Memories
Asynchronous Design
Exploration & Estimation
Adaptive Control

Architecture, IC Design & Control for Digital SoCs
Research topics: Many-core architecture, low-power, System-on-Chip, 3D stacking

ABSTRACT: P2012 is an area- and power-efficient many-core computing accelerator based on multiple processor clusters implemented with independent power and clock domains, enabling aggressive fine-grained power, reliability and variability management. Clusters are connected via a high-performance fully-asynchronous Network-on-Chip (NoC) and feature up to 16 processors. The SoC is being implemented in STMicroelectronics’ low-power 28nm CMOS process and is 3D stacking ready. Target chip area is below 26mm² for a 4 clusters version.

The Platform 2012 (P2012 [1]) project aims at moving a significant step forward in programmable accelerator architectures for next generation data-intensive embedded applications such as multimodal sensor fusion, image understanding and mobile augmented reality. P2012 is an area-, power-efficient and process aware many-core computing fabric, and it provides an architectural harness that eases integration of hardwired IPs. P2012 can be described as a Globally Asynchronous Locally Synchronous (GALS) fabric of tiles, called clusters, connected through an asynchronous global NoC [2] (G-AnoC). The P2012 cluster aggregates a multi-core computing engine (ENCore), and a cluster controller (CC). The ENCore cluster can host a number of processors varying from 1 to 16.

Power, thermal and variability management are essential features in computing architectures targeting deep-submicron CMOS implementation. P2012 makes use of several hardware-assisted control loops to reduce design-time margin and to improve energy efficiency. Each cluster has a local clock, generated with a small-size and highly reactive Frequency-Locked-Loop (FLL). Clock speed can be adjusted in a few cycles on a per-cluster basis with no inter-cluster constraints. The fabric interconnect is fully asynchronous, hence no global chip-wide clock distribution is required. Static and dynamic variability are managed through a number of distributed sensors, both direct (critical path monitors, both embedded and replica-based) and indirect (thermal sensors, both absolute and relative). Sensors are accessible through memory-mapped registers clustered in the Clock Variability and Power (CVP) module which controls process, variability and temperature sensors. Hence feedback-based software policies can be implemented for operating point selection.

The first silicon embodiment of P2012 is the flexible SoC depicted in Figure 1. One key innovation in the physical implementation of the SoC is its flexibility in off-chip connectivity. The die can be configured as an “accelerator chiplet” for three-dimensional die-stacking by appropriately setting the static MUXes shown on the right hand side of figure 1. In this 3D mode (denoted 3D AnoC) the fabric interface to host and main memory goes through three 32 bits data-wide asynchronous IO ports driven by micro-buffers and tied to micro-pads for die stacking. In addition (not shown in the figure), power and ground are also delivered through a “vertical plug”. In this configuration the die will be flipped and stacked on top on a host SoC with CPU, peripherals, standard IOs and DRAM interfaces.

A second 2D configuration is supported by the static MUXes. In this mode traditional board-level high-speed interface (denoted 2D SNoC) links the fabric with the external host and main memory. This interface is physically driven through a smaller number of standard IO pads (two 81-pin ports). The 2D configuration allows simple interfacing with on-board FPGA-based hosts.

The SoC is being implemented in STMicroelectronics’ low-power 28nm CMOS process. Target chip area is below 26mm². The power distribution grid of the SoC is designed to handle power delivery in both 3D and 2D configurations. The chip power consumption under heavy workload is upper-bounded at 4W (at 1.1V, 125°C), but its aggressive power management features enables energy-proportional operation up to a few hundreds mW average power.

References:
Enhancing Cache Coherent Architectures with Access Patterns for Embedded Manycore Systems

Research topics: shared memory, coherence protocols, manycores, memory access patterns

J. Marandola (USP), S. Louise, L. Cudennec, J-T Acquaviva, D.A. Bader (GATech)

ABSTRACT: One of the key challenges in advanced micro-architecture is to provide high performance hardware-components that work as application accelerators. In this paper [1], we present a Cache Coherent Architecture that optimizes memory accesses to patterns using both a hardware component and specialized instructions. The high performance hardware-component in our context is aimed at CMP (Chip Multi-Processing) and MPSoC (Multiprocessor System-on-Chip). We also provide a first evaluation of the proposal on a representative embedded benchmark program, which shows that we can achieve over 50% computing speedup and reduce memory throughput by nearly 40%.

Shared memory paradigms are gaining interest to program multicore systems: the main C compilers already embed support for OpenMP. Indeed, such programming concepts allow improving on legacy code to obtain a reasonable and efficient multicore support. But the age of simple multicores is reaching an end: as the number of cores grows, single buses are replaced by Networks-on-Chip (NoCs), distributed memory, and distributed data-paths: bus spying techniques used to ensure cache coherence are no more applicable. With distributed caches and NoCs, the usual MESI (Modified, Exclusive, Shared, Invalid) protocol for cache coherence must be modified to refer to a given (reference) core called Home Node (HN) which tracks the MESI state of a given cache line for the whole chip. But this technique does not scale well, and is not adapted to embedded devices and applications. First, it can be very talkative as seen in Figure 1, and, second, it does not take advantage of regular memory accesses.

Such regular accesses can be represented as memory access patterns and a research effort was engaged which led to a patent deposit [2]. Improving on the baseline protocol which is the state of the art of shared memory mechanisms for manycore systems, a hardware structure and a specific protocol was designed specifically to handle the pattern based access. An example comparing the same series of data accesses for both protocols can be seen in Figure 2.

Even for such a simple pattern with only 3 elements (the difference grows linearly with the size of the pattern), the number of messages is reduced, and a speculative prefetch is done: once the first element of the pattern is detected, the remaining parts of the pattern is fetched and updated without waiting for any other memory access. Hence, future accesses are automatically prefetched and ready for use, reducing both throughput and memory latency.

A first real-size evaluation of the supposed advantage was done on a simple simulation instrumented with a in-house modified version of the pinatrace Pintool memory analyzer, from Intel’s Pin framework. We showed on that on a two-pass image filter that was chosen for it stresses memory accesses, we obtained a reduction of 37% of message throughput and an acceleration of the application by more than 50% with regards to the baseline protocol alone.

Hence, this protocol, taking advantage of regular memory accesses (patterns), was validated on a program representative of embedded applications. The results show that such an apparatus significantly reduce message and memory throughput and accelerate applications. Such breakthrough can be vital for the future of manycore systems, their programmability and their performance.

References:
Adaptive Stackable 3D Cache Architecture for Manycores

Research topics: 3D, cache, NUCA, manycore

E. Guthmuller, I. Miro-Panades and A. Greiner (UMPC/LIP6)

ABSTRACT: With the emergence of manycore architectures, the need of on-chip memories such as caches grows faster than the number of cores. Moreover, the bandwidth to off-chip memories is saturating. Big memory caches can alleviate the pressure to off-chip accesses. We have designed an adaptive 3D cache architecture taking advantage of dense vertical connections in stacked chips. We also propose a dynamically adaptive mechanism to optimize the use of the aforementioned 3D cache architecture according to the workload needs. We show that our approach can lead to a 50% reduction of both external memory accesses and application execution time.

3D technologies allow the placement of caches on top of processors. This greatly simplifies the circuit floorplan. A big NUCA (Non-Uniform Cache Access) cache can then be used without sacrificing access latency and bandwidth to the cache, and without introducing much distance between the processors.

With the high density of TSVs of recent 3D technologies, very large vertical interconnections can provide a high vertical bandwidth to the distributed 3D cache architecture. Such distributed cache architectures can be built with a large number of vertical access ports as shown in Figure 2. We can even imagine having one access port to the 3D cache per processing unit.

The physical view of the 3D cache architecture that we have proposed in [1] is depicted in Fig 1. This architecture integrates processing units, cache tiles, cache controllers and external memory controllers interconnected by NoCs. Processing units (processors or dedicated hardware accelerators) send requests to cache controllers mapped in the global memory space. The cache controllers then dispatch these requests to cache tiles through a 3D NoC. Cache tiles process requests and, in case of MISS, they transmit those requests to external memory controllers through a 3D ExtMem NoC. In our approach, the processing units, the cache controllers and the external memory controller are placed in the bottom tier while the cache tiles are placed in the top tiers. This organization allows us to build a modular stackable architecture.

The key point in this proposal is adaptability: As shown in Fig. 2, the 3D cache can be statically configured by the Operating System, to allocate a larger private cache quantity to a given application. Moreover, the operating system can also decide to share a given cache tile between one or several applications running in various memory segments to reduce the overall MISS rate while losing the exclusivity of access to this cache tile. In this case, the highly accessed memory segments will occupy a larger storage capacity in the 3D cache. By allowing the OS to control the cache resource allocation, we expect to increase the performances: we can either reduce the MISS rate for a chosen application or for the overall system. In this later case, we also expect to reduce the overall bandwidth requirements to the external memory.

SystemC and RTL models have been written to validate the proposed cache architecture and evaluate performances under intensive workloads. Our experiments show that, when running unbalanced workloads, the tile sharing mechanism can reduce by up to 50% both execution time of the most memory intensive application and the overall traffic to the external memory on a 64 cores manycore architecture with 16 MB of 3D cache per tier. With no sharing, the less memory intensive application is not penalized and the cache provides a better quality of service to this application.

We have synthesized our design in CMOS STMicroelectronics 65nm Low Power node. Each tile is a 1 MB cache with 16 ways and 512 sets, and it includes 564 TSVs to stack up to 4 tiers of 3D cache on top of the manycore. The total area of a 1 MB cache tile using 5 µm wide TSVs and SRAM memories is 6 mm2. The useful area (the data of the cache) is close to 90% of the total tile area, demonstrating the low cost of our approach.

References:

Design-for-Test and Fault Tolerant Architecture for a 3D NoC TSV-based Infrastructure

ABSTRACT: 3D stacking is seen as one of the most interesting technologies for System-on-Chip (SoC) developments. However, 3D technologies using Through Silicon Vias (TSV) have not yet proved their viability for being deployed in large-range of products. One of the main challenges of 3D TSV-based design is regarding testability and reliability. For testability purposes, 3D stacking requires that each die be individually tested before assembly to identify the known-good-die (KGD), and the overall test & repair control is achieved using the JTAG protocol and associated JTAG SWITCH to circulate the tdi/tdo scan chain from/to the above/below die; if a die is not detected, the tdi/tdo muxes are controlled to transport the scan chain in the local TAP controller. When testing a single die, the JTAG SWITCH behaves as a bypass between the external JTAG port (tester connection) and the local TAP controller.

For reliability purposes, since 3D technology does not reach yet a mature production level with predictable yield, it is required to offer a certain level of fault tolerance. We introduce a fault-tolerant architecture using spare TSVs, whose test and repair mechanism is fully integrated in the 3D DFT architecture. For die-to-die communication, we introduce about 12% spare TSVs (1 additional TSV every 8 connection), which are controlled by simplemuxes/ demuxes, this repair scheme allows one TSV failure per group of 8 signals.

Figure 1 : 3D DFT Architecture, based on a JTAG Switch

The newly introduced JTAG SWITCH [3] allows to automatically transport the JTAG signals, and to have the tdi/tdo scan chain circulates in the 3D stack, according to the die position in the stack. To achieve this, an automatic detection of adjacent die is performed using some pull-down cells by adding some die-to-die detection signals forced to logic 1. If a die is detected, the tdi/tdo muxes are controlled to transport the scan chain from/to the above/below die; if a die is not detected, the tdi/tdo muxes are controlled to transport the scan chain in the local TAP controller. When testing a single die, the JTAG SWITCH behaves as a bypass between the external JTAG port (tester connection) and the local TAP controller.

Figure 2 : 3D asynchronous NoC with 3D DFT and TSV repair

The proposed 3D DFT and fault tolerance scheme has been applied to a 3D asynchronous NoC [2], which has been encapsulated by two layers (Figure 2) : firstly a JTAG Test Wrapper (in yellow) in order to test all individual TSV connections, and by a Fault Tolerance Wrapper (in pink) in order to control and repair using the spare TSVs. The overall test & repair control is achieved using the JTAG protocol and associated JTAG SWITCH to circulate the tdi/tdo scan chain within the 3D circuit stack.

References:
In FDSOI technology, $V_T$ is primarily set by the metal-gate (MG) stack work function. UTBB-FDSOI technology offers additional flexibility by setting the BP doping type underneath the BOX, either n or p (Fig. 1). Combining BP and twin-MG process integration allows getting at least 3 distinct $V_T$s (High-VT, Regular-VT and Low-VT). In addition, the BOX dielectric electrically isolates the well from the source and drain of the transistors, which expands the range of possible well bias voltages ($V_{WB}$) and therefore improves the range of possible $V_T$ adjustments, through a high body factor. $V_A$ is only limited by pn-well junctions. FDSOI achieves record low $V_T$ variability because of its immunity to RDF, even with forward body bias (FBB).

Analysis methodology in this paper is based on dynamic margins, obtained through transient simulations. Margins against read stability (RS), read access time (RA) and writeability (WA) failure are assessed using Monte Carlo (MC) calibrated surface-potential based SPICE model.

To strengthen the PU, an Single-PW bitcell architecture is introduced, depicted in Fig. 3. Both PMOS and NMOS transistors are placed over a common P-well, which lowers the threshold of the PMOS transistors. The $V_T$ of PD and PG NMOS transistors does not change, compared to the bulk-like bitcell. By increasing $V_A$, NMOS transistors are forward biased to improve RA and therefore $V_{MIN}$. In this framework, the PW is isolated from the p-substrate by using a deep n-well (DNW) tied to $V_{DDS}$. Thanks to the single common well, $V_A$ can be biased up to (or tied to) $V_{DDS}$, biasing the nMOS transistors in a full forward mode. WA improvement lowers $V_{MIN}$ by 120mV for 64 and 128 bitcell columns, while for 256b RA improvement lowers $V_{MIN}$ by 60mV. RS is also improved by almost one sigma due to the higher strength of the PU, which reinforces the high voltage level on the opposite side of the bitcell during the read stress.

The SPW architecture and $V_B$ biasing improve $V_{MIN}$ at the cost of increased leakage at same supply voltage. For power management in a mobile processor, two back-bias modes can be considered. In active mode, VB would be biased to $V_{DDO}$ to achieve the lowest $V_{MIN}$ and in standby-mode $V_B$ is grounded to minimize the static power, making the SPW bitcell compelling for high capacity cache designs.

This work was carried out in collaboration with the BWRC of UC-Berkeley and STMicroelectronics.

References:
Ultra-Low-Voltage SRAM Design in UTBB-FDSOI technology

Research topics : SRAM, Low power, Low Voltage, Stability

A. Makosiej (ISEP & Leti), O. Thomas, Andrei Vladimirescu (ISEP), Amara Amara (ISEP)

ABSTRACT: In today’s systems-on-a-chip (SOC) the embedded SRAM can often take over 50% of total chip area, which in some cases may lead to the leakage power dominating the overall power consumption. Supply voltage (VDD) scaling is an efficient way to reduce the SRAM leakage but it is limited by the ever increasing parameter variations, which adversely impact the cell stability. Ultra-thin-body and box FDSOI (UTBB-FDSOI) technology emerges as an efficient solution for Low Power SRAM design, due to its strong multi-VT capabilities and low variability. Moreover, the high body factor of this technology allows the optimization of active and standby mode stability, further increasing its potential for Low Power SRAM design.

Minimization of SRAM power consumption is a major concern for modern chip design for a number of reasons: (i) increasing area taken by embedded SRAM - even over 50% of the total area, (ii) increasing individual transistor leakage due to the degradation of electrostatic control of the channel in standard bulk CMOS devices, (iii) large random threshold voltage (Vt) variation- limits voltage scaling range due to stability constraints (AVT=AVT/√(LW), where AVT is the Pelgrom Coefficient) and (iv) contradictory requirements on cell operation conditions for best stability in active and standby [1,2].

UTBB-FDSOI is an attractive solution to reduce the impact of these issues on SRAM power efficiency. Improved electrostatic control of the channel leads to better device properties and hence, lower leakage. Due to an undoped channel the Random Dopant Fluctuations no longer affect variability. In consequence the AVT is significantly reduced and is evaluated at approximately 1.25 mVμm at 22nm as compared to almost 2.5 mVμm for 32nm Low Power bulk CMOS. High body factor of 60-70mV/V and very wide range body bias adjustment are both unique properties of UTBB-FDSOI at sub-28 nm nodes, allowing stability optimization for active and standby modes separately. Finally, due to the possibility of changing the type of the backplane doping, modification of the gate workfunction and the use of either single or two well designs, strong multi-VT capabilities are obtained. The importance of these features for sub-28 nm power efficient SRAM is demonstrated through the analysis of high density SRAM limitations from the point of view of minimum applicable voltages in active (Vmin) and standby (DRV) for 60 years [3]. Both metrics are evaluated using the Static Noise Margin (SNM) approach, under the assumption that read SNM is the limiting factor for active mode stability.

Figure 1 depicts the Vmin in function of AVT for two different Vt cases for the minimum and enlarged cells at 22 nm node. It can be noted, that Vmin decreases significantly towards lower AVT with the slope of over 600mV/1mVμm. Moreover, the optimization of DRV by decreasing AVT causes a simultaneous increase of Vmin by 100mV, indicating the requirement of different Vt sets for Vmin and DRV minimization. Figures 2 and 3 show the DRV and maximum tolerable AVT for Vmin<VDD for various technology nodes. Clearly, the DRV increase with technology node shrinking is significant even for the AVT as low as 1.25 mVμm, in particular for the nominal Vt case. In Fig. 3 it can be observed that in order to maintain Vmin<VDD for minimum sized cell, a 30% decrease of AVT per technology node must be obtained, reaching as low as almost 1.25 mVμm for 22 nm node for nominal and 1 mVμm for DRV optimized case.

UTBB-FDSOI with AVT=1.25 mVμm should therefore allow obtaining Vmin<0.7 V and DRV<0.4 V at 22 nm node under the condition, that a 100mV Vt adjustment through body biasing between active and standby modes is performed and the initial “nominal” Vt values are carefully set using the multi-Vt capabilities.

Figure 1 : Vmin as a function of AVT for 2 different cell sizes and cell Vt ratios; one corresponding to minimum DRV and the other for nominal case with improved read stability (Vmin+100 mV)

Figure 2 : DRV degradation vs. technology node for 2 different Vt sets and AVT=1.25mVμm

Figure 3 : Minimum AVT required to meet the minimum cell sizing requirement from Vmin perspective vs. technology node for 2 different Vt sets; VDDs scaled by technology along LP roadmap (0.65V;0.7V;0.75V;0.8V;0.85V;0.9V, respectively)

This work was performed in collaboration with Andrei Vladimirescu (BWRC, UC Berkeley and ISEP, Paris) and Amara Amara (ISEP, Paris).

References :
A Mixed LPDDR2 Impedance Calibration Technique exploiting 28nm Fully Depleted SOI Back-Biasing

Research topics: LPDDR2, FDSOI, Back-Biasing

ABSTRACT: Signal integrity is a major concern in high-speed interfaces for digital communications. In such interfaces, the impedance of the output driver must be matched to the impedance of the transmission line, in order to avoid signal reflections. Traditionally, a thermometer code is used for adjusting the impedance as function of Process, Voltage and Temperature: segments of the output driver are turned on or off. The drawback is that, during such calibration phases, the interface is unavailable. The back-biasing capability of the Fully-Depleted SOI technology is exploited to compensate for temperature and voltage drifts during circuit operation, while the process deviation remains digitally compensated: this mixed calibration is implemented in a LPDDR2 memory interface.

With the ever increasing communication data rate, the need for achieving high signal quality through transmission lines is mandatory to avoid bit errors. One of the requirements of high-speed interfaces is that their output impedance must be matched to the impedance of the transmission line, in order to avoid signal reflections. Double Data Rate (DDR) interfaces are a good example of a high-speed digital interface – in that case between a microprocessor and a memory. In a conventional LPDDR2 interface, widely used in mobile applications, the impedance of the output driver is matched by using a digital calibration scheme. The LPDDR2 transmitter is made of 7 parallel slices, each one is calibrated to 240Ω whatever the Process-Voltage-Temperature (PVT) conditions, providing a ~34.3Ω total impedance on the output driver to match the transmission line. Each slice consists of N+1 programmable Pull-Up (PU) array, N+1 programmable Pull-Down (PD) array and a shared series resistance (RLIN) for linearity concerns.

During the calibration phase, those PU/PD transistors are turned on or off to adjust the total impedance to the specification. It must be noted that, during this calibration phase, the I/O interface is not functional. The calibration phases occur at only at circuit power-on to compensate for Process variation, but also during circuit operation to cope with voltage and temperature drifts.

We conducted research to avoid DDR interfaces being periodically rendered unavailable because of calibration, by exploiting the Back-Biasing capability of Fully-Depleted SOI (FDSOI) technology.

The threshold voltage of transistors fabricated in Ultra-Thin Box and Body technology can be strongly modulated by applying a bias voltage to the back interface. A mixed analog/digital impedance calibration scheme was thus developed [1]: the digital calibration is kept for the initialization phase, to compensate for Process variation, and the analog calibration is used for coping with temperature and voltage drifts. The mixed calibration scheme is shown in Fig. 1. The back-bias voltage PU_BB is common to all transistors and is generated by a charge pump, driven by a bang-bang controller. Simulations of digital calibration phase followed by analog calibration phase have been carried out. The obtained results, in all PVT corners of the PU versus the PD impedance, are depicted in Fig. 2. It shows that mixed analog/digital impedance calibration fulfills perfectly the ±15% LPDDR2 impedance specification. The PU impedance is within ±6% against ±9% for the PD impedance. Less accuracy can be observed for the PD impedance compared to the PU one. This can be explained by the fact that PU calibration is based on an external accurate resistance, whereas PD calibration is based on a PU replica. Therefore, the possible inaccuracy coming from the PU array is transmitted to the PD array.

References:
Fault Tolerant Asynchronous Design, Design Flow & Application to Network-on-Chip

Research topics: Asynchronous design, SEE, Fault tolerance

J. Pontes, P. Vivet, N. Calazans (PUCRS)

ABSTRACT: In advanced CMOS technology, Single Event Effects (SEEs) due to high energy particles may cause different types of electrical effects when crossing silicon: from small delay variations, to bit flips, until permanent damage. Asynchronous QDI circuits are immune to delay variations but are sensitive to bit flips as are synchronous circuits. Targeting fault tolerant many-core architectures, we propose a temporal redundancy delay insensitive code for application to asynchronous Network-on-Chip. The proposed TRDIC scheme has been validated using a new SEE digital design flow, for accurate SEE fault injection and simulation, in a 32nm CMOS technology.

In advanced CMOS technology, Single Event Effects (SEEs) due to high energy particles may cause different types of electrical effects when crossing silicon: from small delay variations, to bit flips, until permanent damage. Due to their un-clocked nature, Quasi Delay Insensitive asynchronous circuits are the most immune to any delay variations thanks to the use of Delay Insensitive codes, but can be very sensitive to bit flips since a Single Event Effect may corrupt the asynchronous handshake protocol. QDI asynchronous logic has been extensively studied to design robust and low-power Network-on-Chip interconnects for GALS like many-core architectures [1]. Nevertheless, even if large research has been done on fault tolerant synchronous design, very little research has been carried out on fault tolerant asynchronous logic. In this paper [2], we propose a design technique to mitigate Single Event Effect by adding Temporal Redundancy to Delay Insensitive Codes (TRDIC). This multiple bit fault tolerant design technique is adaptable to any 1-of-N DI codes, and is particularly well suited to asynchronous Networks-on-Chip.

As presented in figure 1, the initial 1-of-m DI code is encoded in a 2-of-m+1 DI code, by adding one bit encoding the concatenation of the current data token with the previous data token value, then the new encoded token is sent through the asynchronous NoC data link, and finally is decoded and corrected in case of error detection, by comparing the received token with the expected previous token. This temporal redundancy scheme is less costly than spatial redundancy such as TMR, and requires only a slight modification of the existing asynchronous NoC link and routers (by adding one bit per data token).

The proposed Temporally Redundant Delay Insensitive codes (TRDIC) have been evaluated using a Single Event Effect digital fault characterization environment (figure 2).

Regarding Design-Flow, 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 fill the gap between accurate spice-level simulation of SEE effects and standard fault simulation injecting fault in digital logic at RTL level, we have proposed and developed an accurate digital design flow, which inject and simulate SEE fault propagation [3]. 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. Existing technology libraries are modified to take into account SEE effect, a SystemC fault simulator performs fault injection/monitoring in accurate back-annotated gate level simulations. This can be applied at system level for any std-cell based design, synchronous or asynchronous.

The TRDIC fault tolerant asynchronous NoC architecture and the associated SEE design flow has been fully exercised in a 32nm CMOS technology. The fault simulation result shows better SEE tolerance (figure 3).

References:
ABSTRACT: Aggressive CMOS technology nodes present increasing variability, which impede the implementation of high-performance large scale synchronous circuits. To overcome this, we developed a performance-oriented implementation flow for QDI asynchronous circuits, which is fully compatible with conventional EDA tools for synchronous designs. Using pseudo-synchronous models of a simple standard-cell library for asynchronous logic, a simple set of pseudo-synchronous timing constraints can be given to industrial EDA tools to benefit from their optimization strategies, during synthesis, place & route. This flow allows achieving significantly better performance and regularity than asynchronous modeling, for faster run times and reduced design effort.

While QDI asynchronous circuits are designed to be insensitive to propagation delays, they cannot rely on a global clock frequency target to define their speed, and most often full-custom implementation is needed to achieve high performance.

To overcome this, our method [1,2] considers each flow-control synchronization cell (C-element) as a pseudo-flip-flop, by splitting the timing arcs between its inputs and outputs in two, taking a reference time instant in the middle as a pseudo-clock triggering edge. As these C-elements are the only ones needing a Reset input to initialize their logic state, it is possible to define the global Reset signal as a pseudo-clock with ideal propagation delays. The original in-out timing arcs are converted to a Reset->out combinational arc at the beginning of a pseudo-synchronous timing path, and several in->Reset setup constraints at the ends of pseudo-synchronous timing paths. By doing so, it becomes possible to break all combinational loops without disabling any timing arc, as shown in Fig.1.

The method derives new timing models for the C-elements using initial Liberty '.lib' models coming from standard cell characterization. From the original timing arcs is derived a pseudo-clock with ideal propagation delays. The original in->out timing arcs are converted to a Reset->out combinational arc at the beginning of a pseudo-synchronous timing path, and several in->Reset setup constraints at the ends of pseudo-synchronous timing paths. By doing so, it becomes possible to break all combinational loops without disabling any timing arc, as shown in Fig.1.

Using a dummy clock constraint on the pseudo-synchronous models, it is possible to perform timing-driven synthesis, placement and routing, with up to 60% improvement in speed, for at most 25% increase in area, as shown on Fig.3. With refined pseudo-synchronous constraints for forward and backward paths, it is even possible to achieve maximum performance with zero slack and only 20% area increase.

Figure 2: Implementation flow using pseudo-synchronous models

Our method was successfully applied to implement SoCs using high-speed asynchronous networks-on-chip in 65nm [3,4] and 28nm [5] with peak performance at up to 1.28GHz.

Figure 3: Pseudo-synchronous quality of results vs dummy period

References:
An Iterative Computational Technique for Performance Evaluation of Networks-on-Chip

Research topics: Networks-on-chip; Performance Evaluation; Formal methods

S. Foroutan (now with TIMA), Y. Thonnart, F. Petrot (TIMA)

ABSTRACT: This work introduces a novel analytical method that can be used in the design of best effort wormhole Networks-on-Chip (NoC) for the purpose of performance evaluation and thus optimization loop and design space exploration. The method is based on a router delay model that computes average router latency and other delay components (such as port acquisition and link transfer delays) of a best effort router. The router model is used iteratively to deal with the direct contentions that occur in a router (i.e. iterative technique), in a recursive algorithm (i.e. dependency tree) to deal with indirect contentions and back pressure impact that happens in sequences of routers.

Due to the distributed and complex nature of Networks-on-Chip (NoC) in terms of topology, wire size, routing algorithm, etc, the performance of a NoC-based infrastructure is difficult to predict. Therefore, one of the important phases in the NoC design flow is performance evaluation which is to extract performance metrics in order to verify whether a specific instance from the NoC design space satisfies the requirements of the entire system. In this sense, reducing the time to obtain the NoC performance and consequently speeding-up the design space exploration, is one of the keys that can considerably reduce the design-flow time and cost.

Path latency is counted from the moment a tagged packet header arrives to the input port of a source router until the moment it gets out of the path destination router (Fig. 1). It is the sum of the routers average latencies. Due to resource sharing in Best Effort NoCs, the tagged packet may have contention with disrupting packets coming from other flows. Direct contention (reciprocal impact on average latency that different flows produce on each other) happens in one router and causes a cyclic dependency in the computation of latencies of different flows incoming to the router, which we solve by an iterative technique.

Indirect contention happens in a sequence of routers. In a BE wormhole network a chain of packets with different destinations may stay blocked one after the other over a sequence of routers. This means that the latency of each router is a function of contention and thus of the latency of its following routers (downstream routers in the sequence). To deal with this acyclic dependency, we build a dependency tree (Fig. 2), and then recursively compute router latencies from the leaves of the tree backward to its root.

Using a Poisson analytical traffic model of an application, and from the two-step method using recursively indirect and direct contention, we are able to obtain <5% error on latency up to 80% of the saturation point in seconds of runtime.

<table>
<thead>
<tr>
<th>NoC (2D-mesh)</th>
<th>Path (NoC Diameter)</th>
<th>Simulation Runtime</th>
<th>Analytical Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>4x4</td>
<td>r_{TL} \rightarrow r_{TL}</td>
<td>9393 s (~3 hours)</td>
<td>1 s</td>
</tr>
<tr>
<td>5x5</td>
<td>r_{TL} \rightarrow r_{TL}</td>
<td>16210 s (~5 hours)</td>
<td>2 s</td>
</tr>
<tr>
<td>6x6</td>
<td>r_{TL} \rightarrow r_{TL}</td>
<td>34415 s (~10 hours)</td>
<td>6 s</td>
</tr>
<tr>
<td>10x10</td>
<td>r_{SO} \rightarrow r_{SO}</td>
<td>More than 48 hours</td>
<td>46 s</td>
</tr>
</tbody>
</table>

References:
The emergence of new embedded applications for telecom, automotive, digital television and multimedia applications has fueled the demand for architectures with higher performances, and better chip area and power efficiency. These applications are usually computation-intensive, which prevents them from being executed by general-purpose processors. In addition, architectures must be able to simultaneously manage concurrent information flows; and they must all be efficiently dispatched and processed. This is only feasible in a multithreaded execution environment. Designers are thus showing interest in System-on-Chip (SoC) paradigms composed of multiple computation resources connected through networks that are highly efficient in terms of latency and bandwidth. The resulting new trend in architectural design is the MultiProcessor SoC (MPSoC).

To bring performance increase on such systems, applications need to be parallelized. One possible approach to parallelize an application is to pipeline its execution. This programming and execution model suits well with data-oriented applications that consider a continuous flow of data. However embedded computation-intensive applications become highly data-dependent and their execution time depends on their input data. As a result, static allocation is non-optimal in such systems and pushes forward the need for efficient online control.

SESAM [2,3] is a tool that was designed to help the design of new asymmetric MPSoC architectures. This tool allows the exploration of MPSoC architectures and the evaluation of many different features (effective performance, used bandwidth, system overheads...). In this paper [1], we associate the SESAM environment to a semi-automatic code generation workflow using Par4All. For the first time, two exploration tools, one for the architecture, one for the task code generation of dataflow applications, are associated to create a complete exploration environment for embedded systems. Fig. 1 shows how SESAM and Par4All tools interact to bring this fully integrated exploration environment.

To validate this approach the exploration of the implementation of a radio sensing application on a complete asymmetric MPSoC architecture was conducted. Various parameters of both the application and the architecture were studied to find efficient trade-offs for performance and silicon efficiency. For instance we analyzed the impact of the number of pipeline tasks and the number of processing resources on execution speed (Fig.2), the memory usage or the control overhead, with a limited application parallelization development cost. Thanks to the association of our tools electronic system designers can really tune both the application and the architecture to bring higher performance to the end-user.

References:
As silicon industry is moving towards smaller and smaller critical dimensions, controlling device parameters during fabrication is becoming a great challenge. The variations in channel length, width, oxide thickness and channel doping profiles result in a large variation of the threshold voltage. As the leakage components in a device depend on the transistor geometry and threshold voltage, statistical variation of those parameters leads to a significant spread of the total leakage. This increased variability in advanced CMOS technologies is playing an increased role in determining the total leakage of a chip. This has accentuated the need to statistically account for leakage variations during the design cycle. Designing for worst case leakage may cause excessive guard-banding, resulting in lower performances. Moreover, underestimating leakage variations can reduce the yield, as dies, violating the product leakage requirements, should be discarded.

In this work we propose a solution to estimate the leakage current of large circuits by modeling each gate for every input vector. This work is based on pre-characterizing library cells, and storing data like leakage mean, variance and correlations in tables. Monte Carlo simulations are used as a reference for this work. In this study leakage estimation is performed on technology mapped ISCAS circuits, which consist of fixed sets of cells. In order to enable a fast and efficient estimation, some information is pre-computed for each cell of the library and stored in look-up tables. We also used correlation matrix created for leakage estimations.

Look Up Table approach involves the characterization of the mean ($\mu$), variance ($\sigma^2$), and correlation ($\rho$) of leakage of each cell for each input state. To build each cell table (leakage mean and variance for each input state), we performed 10,000 simulations of Monte Carlo Process Variations (MC-PV) at temperature ($T=125^\circ C$). We explored this approach for the C3540 ISCAS 85 benchmark circuit with a detailed comparison between Monte Carlo and the LUT approach as shown in Figure 1. Variance calculated from LUT approach gives accurate results as we are considering the input state of each cell. Detailed comparison for C3540 is shown in figure 2, Q-Q plot, which compares Monte Carlo quantiles on horizontal axis with LUT quantiles on the vertical axis for ISCAS C3540 circuit. The linearity of the points shows that LUT fits the MC well. From Q-Q plot we can see that blue points are aligned with red points. 99.9% of data are accurately mapped and only 0.1% of data is overestimated by LUT approach, using the cell correlation coefficients.

The major advantage of this approach is to limit the size of the simulations. For ISCAS C3540, Monte Carlo Simulation time was approximately one hour and leakage computation time by LUT approach, with cells correlations, was few seconds for a maximum relative error of 0.4% for correlation factor for bigger circuits.
Workload Impact on HCI-BTI Induced Timing Variations in Processor Microarchitectures

Research topics: processor microarchitecture, simulation, HCI, BTI

ABSTRACT: Hot Carrier Injection (HCI) and Bias Temperature Instability (BTI) failures become the major detractors in the leading-edge MOS technology. High-performance chips will suffer from internal timing shifts and logic errors. A software framework is proposed to analyze the impact of processor workload on HCI-BTI, early in the design cycles. The sensitive paths to the degradation mechanisms, the variations of slack times and the most aggressive instructions are extracted. This work is the first step of a global methodology that aims at enabling design space exploration of multi-processors for reliability in 28-nm and below.

Die shrinking under 32-nm combined with the non-ideal scaling of voltage increases the probability of transistors to encounter Hot Carrier Injection (HCI) and Bias Temperature Instability (BTI). These degradation mechanisms result in a shift of Vth, leading to a loss of performance. The damage rate due to HCI cumulates during transitions between two logic states, while the one due to BTI cumulates when transistor gates have a positive (PBTI) or negative (NBTI) bias voltage. Static Timing Analysis tools (STA) aid to achieve a preliminary sign-off of the chip path timings.

Recent works proposed degradation models of standard cells able to estimate the timing shifts in logic cells during the first design cycles and hence, the variation of the maximum frequency. The analysis is performed by applying random input vectors, whose generation is based on the hardware implementation (e.g., automatic test pattern generator). This approach leads to an over-estimation of the frequency shift. Here, a simulation framework is developed to analyze the impact of the instruction set (RISC) on the timings of a processor microarchitecture netlist under HCI condition [1] (BTI is under investigation). Fig. 1 shows the internal organization of our framework that includes proprietary and third-party tools. RTL design is first synthesized. Then, the applications are simulated in the netlist and bit toggling activity is extracted. Finally, the degradation analysis is performed: aged slack times and sensitive paths to HCI are extracted. A RISC processor named AntX (developed by CEA LIST) and designed in 40nm TSMC technology was investigated (freq=200MHz). Fig. 2 shows the aged slack time of the 70 longest paths vs. various workloads formed of 7 benchmarks and Worst case (all bits always toggle) scenarios. The results are obtained with an identical simulation time (17µs) and an accelerated degradation condition. Path ranking is done under fresh condition. For all applications, the most sensitive path to HCI differs to the critical path. For all paths, the shift varies according to the workload. The worst case scenario (blue line) leads to an over-estimation of the biggest variation (up to x4). The most aggressive instruction is the Shift operation. The execution of instruction pairs, including this instruction and various operands, can aid to tune the design guard band to a more realistic value.

This work is the first step of a global methodology that aims at enabling the design space exploration of multi-processors for reliability in 28-nm and below. Our current work is focused on the development of an enhanced SystemC/TLM based processor simulator with the capability to predict timing shifts at instruction level, hence ready to be integrated in a multi-processor simulator such as described in [2].

References:
Voltage and Temperature Estimation Using Statistical Tests for Variability Mitigation and Power Efficiency

Research topics : Variability, Power management, MPSoC, AVFS

L. Vincent, E. Beigné, S. Lesecq, P. Maurine (LIRMM)

ABSTRACT: Power efficiency of embedded systems has become a tremendous challenge in the context of limited power budget and computational performance constraints. Recent evolutions of Multi-Processor System-on-Chip (MPSoC) architectures provide a way to distribute the workload over several cores while the power consumption can be highly decreased by the application of so-called Dynamic Voltage and Frequency Scaling at fine grain.

The use of advanced technologies implies more variations during the manufacturing process (P) that affect the characteristics and performances of similar chips on the same wafer but also between cores on the same chip. The voltage (V) and temperature (T) variations are environmental variations that also affect the performances of the chip. PVT-variability also affects the power consumption. To reach high performance targets, the traditional worst case approach leads to an increase in the design margins while sacrificing the power efficiency. As VT-variability is local and dynamic, the architecture has to be adaptive (Fig. 1), i.e. the VT variations have to be monitored and mitigated on-chip and at run-time.

Power consumption provides huge constraints on the development of embedded systems because of limited power budget associated with thermal issues. Today embedded applications require even more computational performances. The development of Multi-Processor System-on-Chip (MPSoC) architectures provides a way to distribute the workload over several cores while the power consumption can be highly decreased by the application of so-called Dynamic Voltage and Frequency Scaling at fine grain.

The method proposed is based on the comparison of the current set of frequencies with other ones stored in a model database. The models are acquired during a calibration phase when known (V,T) states are applied to the chip.

At run-time, the comparison between the current measurements and each model is performed using a non-parametric goodness-of-fit hypothesis test. This test measures the discrepancy between the measurements and the model. Fig. 2 presents one result of the estimation procedure. The green circle corresponds to the real (V,T) conditions applied to the circuit while the black cross is the estimated (V,T) state. Each model tested is depicted by a dot whose color corresponds to the discrepancy between the model tested and the measurements acquired from the Multiprobe. Dark red (resp. dark blue) dots correspond to a high (resp. low) similarity between the model and the measurements.

Experiments have shown a mean estimation accuracy of (5mV, 7.5°C) on the estimation of V and T.

In this context we propose an information extraction method (green box on Fig. 1), associated to a fully digital sensor (black squares) to estimate the current local VT state within a power domain. Contrarily to existing voltage or temperature sensors, our sensor [1], named Multiprobe, is a digital one, based on a set of 7 different Ring Oscillators. As it is not possible to infer directly the V and T values from the set of measured frequencies, our method fuses the measurements in order to extract the Voltage V and Temperature T estimates from the Multiprobe measurements. The method proposed is based on the comparison of the current set of frequencies with other ones stored in a model database. The models are acquired during a calibration phase when known (V,T) states are applied to the chip.

At run-time, the comparison between the current measurements and each model is performed using a non-parametric goodness-of-fit hypothesis test. This test measures the discrepancy between the measurements and the model. Fig. 2 presents one result of the estimation procedure. The green circle corresponds to the real (V,T) conditions applied to the circuit while the black cross is the estimated (V,T) state. Each model tested is depicted by a dot whose color corresponds to the discrepancy between the model tested and the measurements acquired from the Multiprobe. Dark red (resp. dark blue) dots correspond to a high (resp. low) similarity between the model and the measurements.

Experiments have shown a mean estimation accuracy of (5mV, 7.5°C) on the estimation of V and T.

We are currently developing strategies to adapt at run-time the settings of the so-called voltage and frequency actuators to reach the most energy efficient operating point.

References : 
[6] S. Lesecq, L. Vincent, E. Beigné, Ph. Maurine, How state estimation in integrated circuits based on statistical tests can be used to fine-tune the control of the voltage and frequency actuators in the power management framework", VARI’2012, Nice, France.
Power Mode Selection in Embedded Systems with Performance Constraints

Research topics: System-on-Chip, Power Management

Y. Akgul, D. Puschini, S. Lesecq, I. Miro-Panades, P. Benoit (LIRMM), L. Torres (LIRMM), E. Beigné

ABSTRACT: Mobile computing platforms must provide ever increasing performances under stringent power consumption constraints. Dynamic Voltage and Frequency Scaling (DVFS) techniques allow to reduce power consumption by providing just enough power to the chip to finish the task before its deadline. DVFS is usually achieved by setting the supply voltage and the clock frequency to predefined values (so-called “Power Modes”) during given durations that depend on the task to be run and on its deadline. Here, the problem of power management is recast as a linear programming one and the time spent in each one of the $N$ power modes is obtained with a Simplex algorithm solution. Results for 3 power modes exemplify the proposed approach.

Moreover, we have shown that for a given range of workload, the 3rd power mode must be carefully placed in order to maximize the power consumption gain. Figure 2 shows such a situation. If the 3rd PM is properly positioned between PM1 and PM2, then the consumption gain for all workloads between 63% and 93% is higher than 15%.

The case with $N > 3$ PMs must be deeply studied in order to express the optimum solution in an analytical form, or at least provide a sub-optimal analytical solution not far from the optimal one. In this way, the solution of the optimization problem, solved in this work with the Simplex algorithm, might be reached through simple computations, possibly implemented in hardware. In our future work, the switching time and the extra power consumed during the transition between PMs will be as well taken into account.

In [1] an analytical solution was proposed for 3 PMs. We have proved that the choice of the PMs $(P_i, F_i)$ and $(P_{i+1}, F_{i+1})$ to run a given task so that the mean frequency $\bar{F}$ satisfies $\bar{F} \in [F_{i+1}, F_i]$ does not always provide an optimal solution from a power consumption point of view. Our results show that for a given distribution of the power modes, the maximum power consumption gain is reached for a particular workload, see Figure 2.

Moreover, we have shown that for a given range of workload, the 3rd power mode must be carefully placed in order to maximize the power consumption gain. Figure 2 shows such a situation. If the 3rd PM is properly positioned between PM1 and PM2, then the consumption gain for all workloads between 63% and 93% is higher than 15%.

The case with $N > 3$ PMs must be deeply studied in order to express the optimum solution in an analytical form, or at least provide a sub-optimal analytical solution not far from the optimal one. In this way, the solution of the optimization problem, solved in this work with the Simplex algorithm, might be reached through simple computations, possibly implemented in hardware. In our future work, the switching time and the extra power consumed during the transition between PMs will be as well taken into account.
Event-driven Power Management for Wireless Sensor Nodes

Research topics: Energy and Power Management, Energy harvesting

J.F. Christmann, E. Beigné, C. Condemine, J. Willemin, C. Piguet (CSEM)

ABSTRACT: Wireless Sensor Networks (WSN) need improved power supply architectures and energy management to reach enhanced life span. Advanced power management techniques are presented that leverage energy harvesting to reduce charge/discharge constraints on the battery. While power efficiency is improved thanks to multiple power paths architecture, energy management benefits from energy driven behavior to optimize application scheduling. Event based energy monitoring is proposed to allow the more complex architecture to be aware of its energy state while maintaining low power consumption.

Energy harvesting is a relevant solution to address Wireless Sensor Networks energy self-sufficiency issue. It indeed allows longer battery life to be reached and thus enhances the network life span. Today's nodes embed various sensors such as temperature, pressure, light or acceleration sensors, data processing elements such as DSP or microcontrollers and communication modules which enable wireless links between the nodes. Powering those devices has to be done with optimized energy efficiency. Usually, electrochemical batteries are leveraged as energy sources, but their finite capacitance implies limited energy autonomy. Improved power management architectures consist in harvesting energy from the environment (light, thermal gradients or vibrations) in order to recharge the battery and enhance the node life span [1].

We proposed to improve the power supply efficiency by preventing the energy to flow necessarily through the battery. A direct power path is set to supply power directly from the energy harvesters to the power loads (Fig.1). A capacitor is used to temporarily store energy, sustaining short but strong energy needs from the loads. The battery is thus recharged in case of energy excess and drained off when harvested power is not sufficient to maintain voltage level on the capacitor.

Two power paths are available for loads power supply: the conventional indirect power path consists in recharging the battery with harvested energy and afterwards draining it to fulfill the capacitor and the additional direct power path only includes the temporary energy storage into the capacitor [1]. The direct power path has much higher power efficiency due to fewer voltage conversion stages but is only available when both energy is harvested and required by the loads. Nevertheless, while applicative task period is usually fixed, output mean power is constant. As available environmental energy is changing, harvested power may be too low to maintain the capacitor voltage. The battery is then drained off to satisfy the application constraints. We propose energy driven schemes to be used to adapt the task period to incoming energy levels. In this case, tasks are performed once enough energy is stored into the capacitor. Fig. 2 illustrates battery’s and capacitor’s energy levels in fixed and adaptive period schemes while input power varies.

Although tasks are not precisely scheduled, power supply is optimized thanks to the only use of the direct power path. Moreover, the battery is freed and the energy autonomy is improved. Advanced low power voltage monitoring is mandatory to control the power paths within the architecture. Event driven voltage monitoring is proposed to avoid useless voltage sampling and prevent heavy monitoring energy cost [2].

Adaptive harvesting aware power management is thus implemented and scheduling can be optimized according to incoming power levels. Algorithms are under development to leverage energy driven scheme which allows battery free operations while maintaining realistic applicative scenario. A dedicated power management circuit is under fabrication to demonstrate power management improvements in a whole wireless sensor node.

References:
Architecture & IC Design
For Emerging Technologies

Neuromorphic Circuits
Beyond CMOS RF Devices
RRAM Circuits
Towards cognitive chips: Design of Spiking Neural Network 
Building Blocks

Research topics: Spiking neurons, Neuromorphic system

ABSTRACT: Stringent energy constraints and the increased variability of modern CMOS technologies impose to design energy-efficient and defect-tolerant hardware accelerators. Additionally, the spectrum of applications is widening, with a shift from computing applications to Recognition, Mining and Classification applications, for which conventional architectures are not effective. Work was carried out on Spiking Neural Network, which are promising for their very efficient information encoding: (1) Analog and digital implementations of spiking neurons were compared; (2) A configurable conduction delay was implemented; (3) 3D stacking was analyzed.

Neuromorphic architectures have been proposed in the past two decades, with the aim of emulating biological spiking neurons on dedicated silicon hardware. Such neuromorphic systems were initially developed to model biological systems and thus better understand their underlying functioning. But nowadays, high-performance applications experience a dramatic shift of their own, from scientific computing to Recognition, Mining and Synthesis applications: few approaches are better positioned than neuromorphic architectures to tackle those applications, while providing inherent defect-tolerant and energy-efficient properties.

Spiking neurons are usually considered to be the neuron model with the greatest potential for applications, thanks to their very efficient coding of information. Their model is described in Fig. 1: the input spikes are weighted and summed up in a leaky tank. If the obtained value reaches a given threshold, a spike is generated on the output.

The area and power consumption of the hardware design can have a significant influence on system scalability, whether emulating biology or implementing processing tasks. This is why we compared the analog and digital implementations of the Leaky-Integrate-and-Fire (LIF) neuron, in 65nm Bulk CMOS technology [1]. It shows that the analog version is 20 times more energy efficient and 5 times smaller than the digital one. Although analog circuits do not scale as much as digital circuits at each technology generation, projections show that the analog implementation should keep its area advantage at least to the 22nm node. And at same area, it will still be 3x more energy efficient.

For designing an analog LIF neuron, we can easily exploit CMOS basic electrical properties: Temporal integration can be realized through capacitive integration, and spatial summation through Kirchhoff’s law. Finally, leakage is an intrinsic behavior of microelectronic devices. One of the drawbacks of analog LIF neuron implementation is the rather large capacitance that is necessary for integrating input spikes. This is why we considered exploiting the density of integration offered by the 3D stacking technology, and especially cleverly benefiting from the parasitic capacitance of Through-Silicon Vias (TSV) [2]. In Fig. 2 are shown different partitioning strategies of a neuron in 3D: schemes (c) and (d) present the most benefits, respectively reducing the area by a factor of 2 and increasing the connectivity by a factor of 8.

Finally, the conduction delay in neural systems has been proven to play an important role in processing neural information. In hardware spiking neural networks (SNN), emulating conduction delays consists of intercepting and buffering spikes for a certain amount of time during their transfer. The complexity of the conduction delay implementation increases with high spiking rates; it implies (1) storing a large number of spikes into memory cells and (2) conserving the required time resolution while processing the delays. The goal of our research was to find a cost-efficient design that supports high firing rates while maintaining good temporal accuracy. We show that it can be achieved using a mixed counter-register implementation which provides a good area/accuracy tradeoff for a broad range of hardware spiking neural networks [3].

The size of the delay circuit increases with the time granularity (temporal accuracy).

References:

Research topics: neuromorphic system, phase-change materials

ABSTRACT: We introduce a novel energy-efficient methodology “2-PCM Synapse” to use phase-change memory (PCM) as synapses in large-scale neuromorphic systems. Our spiking neural network architecture exploits the gradual crystallization behavior of PCM devices for emulating both synaptic potentiation and synaptic depression. The system, comprising about 2 million synapses, directly learns from event-based dynamic vision sensors. When tested with real-life data, it is able to extract complex and overlapping temporally correlated features such as car trajectories on a freeway. The synaptic programming power consumption of the system during learning is estimated and could be as low as 100 nW for scaled down PCM technology.

Phase Change Memory (PCM) devices have been proposed to emulate biologically inspired features of synaptic functionality that are essential for realizing neuromorphic hardware. Among the different types of emulated synaptic features, Spike-Timing-Dependent Plasticity (STDP) has gained a lot of significance recently. STDP is widely believed to be a foundation of learning mechanisms inside the brain.

When the number of neurons and synapses in a neuromorphic system featuring STDP grows large, its implementation on classical computer architecture quickly becomes a severe demonstration of the Von Neumann bottleneck. This is a major reason motivating the research for new neuromorphic memory architectures that could allow in-situ, instantaneous and fully parallel, synaptic-weight updates. From a technological perspective PCM is a good candidate for neuromorphic applications because of CMOS compatibility, high scalability, strong endurance and good retention characteristics.

One of the main limitations of using a single PCM device as a synapse is the implementation of Long-Term Depression (LTD), which is not progressive with amorphization by using invariant or identical pulses. To overcome these issues, we propose to implement both Long-Term Potentiation (LTP) and LTD using crystalization, with two PCM devices constituting one synapse (Fig.1). The two devices have an opposite contribution to the neuron integration.

This novel low power architecture called ‘2-PCM Synapse’ is aimed for emulation of synaptic functions in large scale neural networks [1,2]. Using this architecture we designed a fully connected, feed-forward spiking neural network (SNN) and implemented a simplified form of the biological STDP learning rule. We show a real world application of extracting complex patterns from recorded video data.

We have demonstrated that Phase Change Memory devices can be used to elaborate large scale synapse-like arrays for neuromorphic systems. A two layer spiking neural network with about 2 million synapses and 4 million PCM devices has been simulated (see Fig.2), showing a complex visual pattern extraction with an average detection rate of 92%, and a synaptic power consumption of 112 µW during learning. The extrapolated power consumption for the most recent state of the art devices, if used for the same test case, could be as low as 100 nW. The low spiking frequency in this type of neural network is remarkable, considering the complex detection task involved, and is a good indicator of the scalability and potentially high efficiency of the association of dynamic vision sensors and spiking neural network compared to the classical synchronous frame-by-frame motion analysis.

References:
ABSTRACT: We present an original demonstration of an associative learning neural network inspired by the famous Pavlov's dogs experiment. A single nanoparticle organic memory field effect transistor (NOMFET) is used to implement each synapse. We show how the physical properties of this dynamic memristive device can be used to perform low-power write operations for the learning and implement short-term association using temporal coding and spike-timing-dependent plasticity–based learning. An electronic circuit was built to validate the proposed learning scheme with packaged devices, with good reproducibility despite the complex synaptic-like dynamics of the NOMFET in pulse regime.

We propose an original scheme using Nano-particle Organic Memory Field Effect Transistors (NOMFETs) to implement dynamic associative learning and demonstrate it by interfacing NOMFETs to a CMOS discrete circuit [1]. We show how the unique synaptic properties of the NOMFET can be used directly at device level to implement what we call a dynamic associative memory, where the association is only retained as long as there is a minimal activity at its input.

In classical conditioning, associative learning involves repeatedly pairing an unconditioned stimulus, which always triggers a reflexive response, with a neutral stimulus, which normally triggers no response. After conditioning, a response can be triggered for both the unconditioned stimulus and the neutral stimulus. This concept goes back to Pavlov's experiments in the early 1900s. He showed how a neutral stimulus - like the ring of a bell - could be associated to the sight of food and trigger the salivation of his dogs (Fig.1). Associative memory is now a key concept in the learning and adaptability processes of the brain.

However, the lack of an efficient implementation of artificial synapses for associative learning neural networks has greatly impeded the use of associative memory as a general-purpose type of memory or learning tool. Synapses implemented with the most current CMOS technology are still several orders of magnitude behind their biological counterparts in terms of area and power consumption, which can be roughly estimated to $10^{12}$ synapses/cm$^2$ and 1-10 fJ/spike if one considers an average firing rate of 1-10 Hz, giving a power consumption of the human brain on the order of 10 W.

It was shown in previous work that the NOMFET [2] could be seen as a memristive device [3], by modulating the conductivity of its organic semi-conducting channel through the charging of nano-particles embedded into the channel. It is volatile and has a retention time of typically 10 to 1000 s. With these physical properties, the NOMFET can mimic / reproduce many behaviors of a dynamic synapse.

Associative memory is a fundamental computing block in the brain and is implemented extremely efficiently in biological neural networks. An efficient and scalable implementation of associative memory would certainly benefit many applications, especially in the area of natural data processing. We demonstrated experimentally an elementary associative memory, which uses only one NOMFET memristive nano-device per synapse (Fig.2). It exhibits dynamical behaviors closer to biology than any other known memristive device.

References:
Spin-Torque Oscillator Modelling for Network Synchronization

Research topics: spintronics, oscillator locking, synchronization, LTI model

M. Zarudniev, P. Villard, E. Colinet, U. Ebels (Spintec), M. Quinsat (Spintec), G. Scorletti (AMPERE Lab)

ABSTRACT: The spin-torque oscillator (STO) is a new device based on thin-film magnetic effects. Its high compactness and wide frequency tuning range are interesting features for multi-standard radio-frequency applications. However, the output signal provided by a single STO exhibits low power and poor phase noise. Interconnecting several STOs within a synchronized network could circumvent these drawbacks. Since the available physical model is too complex for a straightforward study of an STO network, the first step consists in simplifying it to derive a linear time-invariant (LTI) model providing the phase as output signal. Such an LTI model is required to take benefit from the powerful analysis and synthesis methods developed in control theory.

Thin-film magnetism studies in the last decades have opened a new research field, referred as spintronics, where not only the charge but also the spin of electrons are exploited. The properties exhibited by such thin films led to the development of new devices with outstanding performances, like for instance the hard disk giant magnetoresistance read-head. More recently, it has been shown that a spin-polarized DC current of sufficient density can give rise to a high frequency oscillation of a thin-film magnetization. This property is exploited in a new device called the spin-torque oscillator (STO), depicted Fig.1, which provides a GHz-range frequency oscillation of a thin-film magnetization. This new model was derived in two steps. Thanks to a first approximation, the LLGS was turned into a 2-dimensional set of non linear equations in polar coordinates like for instance the hard disk giant magnetoresistance read-head. As an example, Fig.2 shows the phase response of the three models to an input current step. The main component is accounted for, which allows the next step, i.e. the study of synchronization within an STO "macro-oscillator" according to a method relying on analysis and synthesis tools used in control theory.

References:
Carbon Nanotube FET Process Variability and Noise Model for Radiofrequency Investigations

Research topics: CNT-FET, RF integrated circuits, process variability

J.L. González, B. Martineau (STMicroelectronics), D. Belot (STMicroelectronics)

ABSTRACT: This work focuses on process variability and noise in carbon nanotube field-effect transistors (CNFET) to obtain a compact model usable for radiofrequency (RF) design and simulations. CNFET figures of merit (FoM) are determined and compared to International Technology Roadmap for Semiconductors (ITRS) requirements on conventional analog silicon-based devices. The developed model is also used to investigate on the impact of manufacturing process variability on the CNFET’s RF-performance and noise behavior in carbon nanotube field-effect transistors (CNFET) to obtain a compact model usable for radiofrequency (RF) design and simulations. CNFET figures of merit (FoM) are determined and compared to International Technology Roadmap for Silicon.

The quest for a substitute device type for today’s Silicon-based CMOS technology has started some years ago. One of the candidates is the carbon nanotube field-effect transistor (CNFET). This device is expected to provide high-speed and low power, and has been widely investigated in the field of digital devices. Their application to RF circuits has deserved less attention. In this work we investigate the properties of this type of device from the point of view of RF figures-of-merit [1]. We do this by considering not just an ideal device composed of a parallel array of equal carbon nanotube, but a real device in which the tube diameters of a single device may vary following statistical distributions, such as those shown in Figure 1. Moreover, the fact that some of the nanotubes of the device may be metallic is also considered. Furthermore, a non-ideal metallic tubes removal process is also considered, as illustrated in Figure 2.b.

The CNFET model developed also incorporates the noise sources (thermal, flicker, shot) that are found in this type of devices. This realistic device model is used to extract meaningful RF figures of merit and to investigate its dependence on the CNFET manufacturing process and device biasing conditions, as illustrated in Figure 2.

The most important results of this study have been published in [2]. We have shown that excellent performance, except of high flicker noise, is achieved for a large-diameter distribution with low standard deviation. An important outcome of the work is a compact CNFET model incorporating noise and variability modeling suitable for RF circuits design.

References:
Bipolar OxRRAM-based Non-volatile SRAM (NV-SRAM) for Information Back-up

Research topics: Resistive RAMs, NV-SRAM, Low leakage, Non-volatile, FDSOI

As technology scales down, power consumption becomes significant in the overall Systems-on-chip (SOCs) power. A potential solution for reducing static power consumption is to use the technique of power gating – switching off the power supply to certain blocks which are not in use. But this technique cannot be used to the memory section having embedded SRAMs covering almost half the area of the chip, as it will result in loss of information. Another potential solution for reducing static leakage is scaling down the VDD of memory blocks during stand-by but its efficiency is limited by the transistors variability.

Hybrid memory circuits combining Static RAMs (SRAMs) and Resistive RAMs (RRAMs) is an effective solution for information storage during a power-down at a lower leakage. The use of Resistive RAMs (RRAMs) with volatile memories like Static Random Access Memory (SRAMs) will not only provide non-volatility but also reduce the static power consumption on a chip where the leakage current contributed by SRAMs dominates the power consumption of a chip. The hybrid memory cell (NV-SRAM) that realizes these features is composed of a typical 6T- SRAM cell based on 22nm FDSOI technology with resistive OxRRAM devices embedded at the data nodes of the SRAM cell. In order not to lose information during a power-down, the logical state of the volatile SRAM memory is stored into the non-volatile resistive memories. To make the circuit resilient to information loss, the following operational sequence must be respected: RESET, STORE, POWER DOWN, RESTORE (as shown in Fig. 1).

The stability study on the NV-SRAM cell when Vt variability of the transistors is taken into consideration reveals that the key yield limiting factor of the NVSRAM cell is the reliable recovery of data from the OxRRAMs to the SRAMs which strongly depends on the resistance ratio (R_OFF/R_ON) of the OxRRAM devices. The STORE operation is always successful as it will set the OxRRAMs resistance to a low value but R_ON will vary under the Vt variation which will change the R_OFF/R_ON in turn will influence the proper restore of SRAMs logical state from OxRRAMs. Fig. 2 depicts that the (R_OFF/R_ON) can be increased by having a long STORE time or increasing the WLP of the SRAM pull-up transistor. But having a long STORE time results in power consumption and WLP size will be limited by the cell area and the write ability of the SRAM. Fig. 3 depicts a worst case mismatch analysis of the hybrid cell during a RESTORE operation. Referring to Fig.3, the maximum achievable R_OFF/R_ON(=4) of NV-SRAM corresponds to n = 2 which is very low from stability viewpoint. To ensure SRAM yield of 5σ (6σ) for reliable recovery, maximisation of the resistance drop, RDROP in the high δR/δt region corresponding to the R_OFF/R_ON (Fig.2) for low STORE time should be the focus. The results on Monte Carlo simulation for recovery operation indicates that even in the worst case for SRAM yield, the resistance ratio should be at least 10(20). Hence, the stability analysis implies that, from the RRAM technology point, the focus should be to maximise the ROFF/RON resistance ratio.

References:
ABSTRACT: “Normally off, instantly on” applications are becoming common in our environment ranging from healthcare to video surveillance. In such a context, Field Programmable Gate Arrays (FPGAs) present a good trade-off between performance and flexibility. However, they consume high static power and can hardly be associated with power gating techniques due to their long context restoring phase. In [1], we propose to integrate non-volatile resistive memories in configuration cells to instantly restore the FPGA context. We then show if the circuit is in ‘ON’ state for less than 42% of time, non-volatile FPGA starts saving energy. Finally, for a typical application with only 1% of time spent in ‘ON’ state, the energy gain reaches 50%.

Many new embedded applications can be characterized as “normally off, instantly on”. These applications share a similar feature: a long idle period followed by a short highly intensive computing phase. Wake-up phase must be maintained as short as possible in order to avoid missing important information. For example, a heart attack can be anticipated by analyzing electrical signals after an abnormal event, but false detections must be avoided due to complex signal monitoring and processing must be finished in a few microseconds after wake-up.

In this context, General Purpose Processors (GPPs) are not efficient as they offer reduced power efficiency and require long boot sequences when putting to ‘ON’ states. Application Specific Integrated Circuits (ASICs) or System-On-Chip (SoCs) are power efficient but are expensive to develop and not flexible enough to address a wide range of application. Field Programmable Gate Arrays (FPGAs) offer a good compromise between flexibility and power efficiency. However, decreasing supply voltages and shrinking feature sizes increase the leakage current of FPGA which becomes the major cause of power consumption during standby mode. In order to improve static consumption, power-gating technique is massively employed. However, FPGA loses all the information contained in SRAM memories when switched off. FPGAs can communicate with external Flash memories to store their context. Restoring a context after a power-off mode is accomplished by serially loading a bitstream in all the configuration cells of the FPGA which is quite long and can take up to hundreds of milliseconds. Thus, current FPGA structures cannot fulfill ON/OFF application constraints.

With VPR5 toolflow evaluation, NVSRAM integration in FPGA results as 7%, 18% and 2% overhead in delay, area and power respectively. Power gating implementation extends the area and critical path by 8% for each metric. Duty cycle of the application has a direct influence on the conserved power levels. Fig. 2 shows that for the on/off application, if the duty cycle is much lower than the average BET (42%), the power gain increases rapidly. For an application where the circuit is active for 1% of the time, power gain reaches 50% on average.

In [1], we propose to integrate Resistive RAM (RRAM) in the FPGA structure to obtain an instant power-on phase and saving power in “normally off, instantly on” applications. We focus on the ON/OFF application using novel NVSRAM memories with bipolar OxRRAM technology [2] in 22nm LETI-FDSOI process. All the volatile SRAM nodes are replaced with NVSRAM, to store and restore bitstream quickly in an ON/OFF cycle. As a result, a total power down state, i.e. zero leakage power consumption, is achieved. When FPGA is switched into sleep mode, the dissipated leakage power is conserved. However, the overhead due to RRAM integration and power switches must be evaluated. Fig. 1 shows a conceptual view of expected power overhead and gain of the SRAM and NVSRAM implementations. Considering these factors, break-even time (BET) is defined as the duty cycle when the energy overhead is equal to the leakage energy as given in Fig.1. Consequently, when the actual duty cycle of the application is smaller than the BET, it is possible to save leakage power and reduce total power consumption.
Embedded Software

Real Time Software
Parallel Software
Middleware for Sensor Networks
Embedded systems have evolved from a single processor architecture to Multi-Processor Systems-on-Chip (MPSoC). Designing embedded software for these increasingly complex and heterogeneous platforms in an efficient manner is becoming a serious challenge. Applications are particularly difficult to program on platforms based on Network-on-Chip (NoC) interconnection, where developers must also define and setup the communications between the units. Studying and optimizing the mapping of micro-code on the different units is particularly difficult and subject to errors. Developers usually need to define fine-grain and coherent configuration files for every node. Thus, moving a task from one IP core to another, impacts not only the communications of this code but also the ones of all other communication sets it communicates with. Design flows for NoC-based SoC is an active research area. One of the most challenging issues is to organize the mapping on the cores while taking into account issues related to communication like deadlock, bandwidth, latencies, and non-functional properties like energy consumption. Figure 1 shows a generic design flow for NoC-based embedded systems.

Figure 2 illustrates a novel integrated exploration workflow more suitable for hardware and/or application mapping exploration. In order to tackle the exploration bottleneck of the previous workflow, it considers an XML-based model that centralizes the system specifications for hardware design and software mapping. This global database allows sharing the information between the different exploration facilities: HW design tool suite, NoC configuration tools and unit-specific compilers. An IP-XACT compliant representation has been chosen in order to match a standard representation for the hardware design.

This standard-based design flow gives a large speedup for design space exploration, from months to weeks, for less than 10% of overhead compared to the previous design model. In addition, it maximizes the possible block reuse among several projects, allowing the integration of the hardware and software designs ensuring the coherence of the descriptions. Thus, it further reduces the time to explore architecture options, either in the hardware design or the mapping of the application. This approach could be used to automatically explore the design space, optimizing both functional and physical behavior of the overall system.

References:
Optimization of Dynamic Memory Allocation and Associative Arrays for Complex Embedded Software

Research topics: embedded systems, control, associative arrays, memory allocation

ABSTRACT: Embedded systems were initially dedicated to execute very specific tasks leveraging highly dedicated hardware. Since the emergence of on-chip manycore architectures and the convergence between general purpose and embedded architectures, embedded systems are now targeting a wider range of applications. Embedded software involves more and more complex runtime control for parallelism and dynamism.

We characterized new embedded control software such as dynamic scheduling, resources management, scripting languages, binary translation and just-in-time compilation, showing that dynamic memory allocation and associative array manipulations represent a significant part of this new control software. We first focused on memory management optimization and then on associative arrays. We observed that memory allocator algorithms are in fact based on associative array manipulation.

Embedded systems often exhibit highly complex memory organizations. Distributed and private memories, absence of address virtualization are frequent burdens of embedded developer. Moreover, with the increasing amount of parallelism, an additional stress is put on dynamic memory allocation. In [1], we propose a flexible memory allocator to handle complex memory organizations of embedded systems. The memory allocator leverages dynamic code generation so that flexibility is not at the expense of performances. We show that combining dynamic code generation and machine learning, can give a 56% speedup on memory allocator’s allocation and release operations (Figure 1).

During our experiments, we showed that associative array manipulation is a significant part of embedded control software (from 25% for dynamic compilation, up to ~80% in resource managers, binary translators or interpreters). Moreover, we also showed that associative array manipulation is at the heart of memory allocators. Thus, we argue that optimization of associative arrays is a crucial point to deliver high-performance embedded control software.

In order to optimize associative array manipulation in embedded control software, we propose a low-level optimization using self-modifying code. In [2], we base our work on Red-Black tree algorithms which are widely used to implement associative arrays (C++ STL, jdmalloc). In order to accelerate Red-Black tree algorithms, we propose to transform tree data structures into executable code. With Red-Black trees encoded as specialized binary code rather than data, we intend to accelerate the tree traversal by taking advantage of the underlying hardware: program cache, processor fetch and decode. Our experiments show that we obtain a gain of 45% with our technique on an ARM Cortex-A9 processor. We also show that we transfer most of the data-cache pressure to the program-cache, motivating future work on dedicated hardware.

References:
A Practical Approach Towards Static DVFS and DPM Scheduling in Real-Time Systems

K. Trabelsi, M.Jan, R.Sirdey

ABSTRACT: Dynamic Voltage and Frequency Scaling (DVFS) and Dynamic Power Management (DPM) are the two main techniques that can be used at the software level to reduce the power consumption of embedded hard real-time systems. In the field of embedded real-time systems, such as automotive or energy distribution, microprocessor manufacturers have recently proposed chips with not only DVFS capability but also a large number of low-power states that can be used to apply DPM strategies. In order to leverage the energy-saving abilities of such microprocessors and to avoid missing deadlines, we formulate the problem using a linear programming approach in which we modelize the available frequencies and low-power states [1].

In the last few years, a lot of works have been done in the field of power-aware computing, particularly on embedded systems containing real-time constraints. At the software level, there are two main strategies to reduce the power consumption of embedded processors: 1) lower their frequency and their voltage 2) set them in low-power states. The first approach is called Dynamic Voltage Frequency Scaling (DVFS), while the second is called Dynamic Power Management (DPM). DVFS policies change the voltage and/or the operating frequency, more generally the speed, of the processing unit on-the-fly and reduce the dynamic energy consumption. In CMOS circuits, power consumption is proportional to the product of the frequency and the square of the supply voltage. DPM policies reduce the (static) energy consumption by essentially turning off some parts of the system. This allows for a drastic decrease in power consumption, but most of the time the DPM energy-saving modes disable processing, by turning off memories, caches or even processing units completely. Leveraging frequencies and energy-saving modes of modern processing units in the same time is one of the challenges of today's embedded systems community. In the context of hard real-time systems, energy-aware policies must be part of the scheduling, and the challenge is to reduce energy consumption while still meeting the deadlines of tasks. Actually, lowering the speed of the processor increases the execution time of a task.

A simple model can be for instance that the worst-case execution time (WCET) of a task scales linearly with the processor frequency. More realistic models have been proposed which divide a task between computationally intensive parts and peripherals accesses. In all cases, the challenge when using DVFS is that this increase of the execution time must not prevent tasks to fulfill their deadline. When the processor is set to a low-power state, tasks can no longer be executed. In this case, the challenge is to awake the processor from a low-power state just in time so that tasks can still fulfill their deadlines. Some of the properties of hard real-time systems can be used to save power: at compile-time, the tasks are thoroughly examined to get their worst-case execution time (WCET), and their occurrence patterns (period and arrival time). This allows the scheduling algorithm to do a first pass of off-line optimization: computing a minimal speed at which the system can run or the low-power states that can be used to achieve a minimal energy consumption while still meeting the deadlines of the tasks. The scheduling policy can also reduce the energy consumption on-line, by taking into account the dynamic behavior of the tasks.

To the best of our knowledge, previous works have focused on static DVFS scheduling, while static DPM has rarely been studied. The availability of microprocessors with advanced DPM and DVFS functionalities motivates the first contribution of this work: an accurate modeling of DPM and DVFS capacities in order to find the optimal static solution interplaying both techniques. We formulate the problem using a linear programming approach in which we modelize the available frequencies and low-power states using binary variables. Then, we study the difference between such an accurate modeling of DPM and DVFS and when transition costs of these techniques are neglected.

We performed some experiments by using the STM32L microcontroller based family. Figure 1 shows the energy consumption of the DPM, DVFS and the hybrid strategies for a frame of 1 ms.

Figure 1 : Energy consumption of the DPM, DVFS and the hybrid strategies for a frame of 1 ms.

To the best of our knowledge, previous works have focused on static DVFS scheduling, while static DPM has rarely been studied. The availability of microprocessors with advanced DPM and DVFS functionalities motivates the first contribution of this work: an accurate modeling of DPM and DVFS capacities in order to find the optimal static solution interplaying both techniques. We formulate the problem using a linear programming approach in which we modelize the available frequencies and low-power states using binary variables. Then, we study the difference between such an accurate modeling of DPM and DVFS and when transition costs of these techniques are neglected.

We performed some experiments by using the STM32L microcontroller based family. Figure 1 shows the energy consumption of the DPM, DVFS and the hybrid strategies for a frame of 1 ms. Our results show that, using the hybrid strategy, mixing DVFS and DPM, is the best solution regardless of workload and considering transition costs or not. We show also that hybrid strategy whittles down transition costs effect on energy saving.

As future work, we would like to generalize the results we obtained from our generic model on different microcontrollers and not only on the STM32L. The solution described in this paper is pessimistic as it assumes the worst case execution time (WCET). We are therefore also interested in adding robustness to our approach by adding bounds on the uncertainty of WCET.

References :
Dynamic Code Generation: Large Spectrum, Many Applications

Research topics: dynamic code generation, compilation, optimization

H.P. Charles, D. Couroussé, Y. Lhuillier, V. Lomuller, A. Carbon

ABSTRACT: Dynamic Code Generation is used in many situations: to enhance portability, adapt binary code to run-time values, use specialized instruction sets, reduce code size. We show our activity in this domain on significant example. We give results on one: the matrix multiplication dynamic library

To generate an executable program, programmers generally use a compiler which transforms source code to binary code (compile time), and during the execution (run-time) the binary program does not change. This is the classical way to produce and execute binary programs, called static compilation scheme.

There is an increasing number of situations where this classical scheme is insufficient, and sometimes counterproductive.

Due to the increasing complexity of memory hierarchy and parallelism levels, program performance is increasingly tied to the characteristics of the data sets. Data size, data alignments and data values have a deep impact on program behavior. Iterative compilation is a technology allowing a binary code to be adapted to a running data set, but the adaptation is only valid for one specific data set, not the arbitrary set that a user can provide. The adaptation of a running code to a given data set parameter is one of our motivations.

Portability issues on embedded systems, such as cell phones and set-top boxes, is another motivation. Many compilation infrastructures are used such as JIT compilation for Android Java based applications, for javascript in browsers for smart phones and set-top boxes, for graphic rendering on Android GPUs, etc. In these situations, binary generation has to be fast, occupy a small memory footprint, generate efficient code, and use power sparingly.

We have developed a tool and experiments that try to tackle these problems. Usage examples are listed in the following items:

ISA dynamic adaptation: we have developed a small code generator embedded in the Scilab mathematical solver that is able to determine at run-time if a code should run either on a CPU or on a GPU. On the GPU side, our code generator is able to dynamically adapt the code according to the matrix size and based on initial benchmarking.

The results of the experimentation are shown in the figures 1 & 2. Figure 1 shows the performance of MAGMA, the reference mathematical library for matrix multiplication on GPUs. For clarity we have only plotted the results superior to 145 GFlops for matrix sizes between 64x64 and 2000x2000. In figure 2 we have plotted the results from the same experiment using our library dynamic adaptor. [1]

JIT hardware acceleration: [3] JIT compilation requires a lot of memory access, which uses hash tables and tree balancing. This impacts the performance because code generation is done at run-time. We have shown in our article that we could mix data and programs in order to accelerate searches in these trees.

References:
Adapting Just-In-Time Compilation to Embedded Systems

Research topics: JIT compilation, embedded systems

ABSTRACT: Just-In-Time (JIT) compilation is today widely employed in many application domains and massively transferred to embedded systems. However, JIT compilation complexity lead to important performance loss for embedded processors due to their lack of mechanisms to manage JIT compilation algorithm irregularities in terms of control and data. Managing these irregularities, associative arrays and dynamic memory allocation still represent 25 % of the LLVM bytecode compiler execution time, despite many existing software optimizations. To reduce their impact on execution time, our ongoing work consists in the proposition of hardware dedicated resources to accelerate them, based on standard libraries and replacing these software optimizations.

Just-In-Time (JIT) compilation has become a major topic for academic and industrial researchers in the last 15 years. JIT compilation technologies consist on executing all, or parts of, compilation stages dynamically during the application execution. The main reasons of this growing interest are the following:
- Increasing dynamism of applications and their workloads
- Increasing interactions between applications
- Increasing portability and security requirements
- Increasing performance requirements

Based on a state-of-the-art analysis, we identify four main technologies using JIT compilation:
- Virtual machines (eg. Java Virtual Machines)
- Dynamic binary translation (eg. Apple Rosetta)
- Multistage dynamic compilation, consisting in deporting compilation phases to runtime (eg. LLVM framework)
- JIT compilation for dynamically-type languages (eg. JavaScript, Python)

In all these technologies, the efficiency of Just-In-Time compilation depends on the ability to compensate its overhead with execution speedups obtained on the generated code. Compilation algorithms are complex and particularly difficult to handle for embedded processors, with important performance loss introduced, as presented in Figure 1.

![Figure 1: Execution time slowdown comparison between conventional and JIT compilation algorithms.](image)

We profiled different JIT compilation algorithms [1], extracted from the highlighted technologies, run them on a x86 processor, and compared their behavior to regular algorithms extracted from miBench benchmarks. Figure 2 shows a common control irregularity compared to regular algorithms with a slight increase of misprediction rates for indirect branches and instruction cache. This highlights a common complexity on control and a high instruction/data ratio contrary to regular algorithms, in which the amount of managed data is far bigger than the amount of instructions.

![Figure 2: Misprediction rates for indirect branches and instruction cache, relative to the total number of instructions (x86).](image)

Table 1 shows a common data irregularity with a slight increase of indirection depths for JIT compilation algorithms (LLVM in our case), highlighting the fact that JIT compilation algorithms are significantly pointer-intensive.

<table>
<thead>
<tr>
<th>Standard</th>
<th>Compression jpeg</th>
<th>Dijkstra</th>
<th>LLVM compiler</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2</td>
<td>192</td>
<td>75541</td>
<td>7733</td>
</tr>
</tbody>
</table>

Table 1: Algorithm indirection depths (x86 instrumented simulator).

These irregularities, already visible on x86, are important issues for embedded processors due to a lack of mechanisms to handle them (especially concerning predictions). To highlight the parts of the code responsible for these irregularities, we profile the LLVM bytecode compiler. Results obtained show that associative array and dynamic memory allocation represent on average 25 % of its execution time [2], despite many existing software optimizations: LLVM developers provide more than 8 specific data-types re-implementing standard data types of the C++ STL library. Our ongoing work deals with the development of hardware dedicated resources, based on standard libraries. We are looking for accelerating associative array management and dynamic memory allocation, replacing existing software optimizations to reduce their impact on execution time.

References:
Parallelism Reduction in Dataflow Oriented Programming Languages for Many-Core Architectures

Research topics: Parallelism, dataflow, compilation, many-core

The growing number of processing cores on a single chip leads to new challenges regarding the efficient programmability of many-core architectures, while staying appealing to regular HPC developers. In such a context, the dataflow programming model can be used to structure the source code, organize tasks into a logical network and enforce some relevant properties for distributed and shared systems like the detection of live deadlocks and buffer overflows. The Sigma-C [1] programming language has been introduced for this purpose. One of its leitmotiv is to let the developer focus on the algorithm side, while taking care, at compilation time, of the entire parallelism tuning aspects. The goal is here to adapt the number of running tasks regarding the number of physical processing elements in order to 1/ leverage the application performance and 2/ minimize the memory footprint, the latter being a prevalent property in embedded systems.

Tuning a dataflow application can be achieved by transparently altering the task connection graph in a way that 1/ the resulting graph meets the given parallelism requirements and 2/ the application semantics are preserved. Some systems have been proposed in the literature, mostly based on simple task fusion, like in Streamit. In [2], the authors propose a language to describe and modify system tasks in charge of the data reorganization.

The application transformations can also be described using graph patterns [3]. A pattern is defined as a parameterized description of a sub-graph that can be matched in an application and replaced by another parameterized sub-graph. One popular example is the split-join cascade pattern as shown in Figure 1. Provided some good properties regarding the consumptions and productions of the system tasks, it is possible to replace this pattern with another one that is built with a different number of split outputs, join inputs and even a different number of tiers. The resulting pattern directly modifies the number of user tasks instantiated within the cascade pattern. Several patterns have been proposed, designed for generic or application-specific use.

A parallelism reduction engine has been implemented within a plugin for the Sigma-C compilation toolchain. This engine applies different pattern substitutions to find a solution that fits the application onto the targeted host. Figure 2 shows the accuracy of the reduction engine. Two applications including some split-join and systolic matrix multiplication patterns are initially built with respectively 1925 and 1867 instances. The reduction engine is thereafter able to reach a number of instances, given as a parameter, with a small error deviation. Other experimentations show that the engine is able to find the best solution for an application made of 6 patterns, by only taking a few seconds when run onto a low-class laptop computer. This demonstrates that the reduction engine can be used within a regular compilation process.

As for ongoing works, a generic reduction approach is currently evaluated, based on the automatic equivalent task merge. The merging process is thereafter applied such as the Sigma-C application throughput constraints remain satisfied, while reducing at most the memory footprint.

References:
A Parallel Simulated Annealing Approach for the Mapping of Large Process Networks

F. Galea, R. Sirdey

ABSTRACT: We propose a parallel simulated annealing approach to solve a dataflow process network mapping problem, where a network of communicating tasks is mapped into a set of processors with limited resource capacities, while minimizing the overall communication bandwidth between processors. The speedups obtained using this approach enable us to solve problems with more than one thousand tasks, on up to 48 processors, in reasonable time. Results have been obtained by taking profit of the specific architecture of a Non-Uniform Memory Access (NUMA) computer.

In the context of compilation for manycore architectures, data flow programming languages such as ΣC [2] allow to design parallel applications which can naturally exploit the massive parallelism of the target architecture. Dataflow applications are made up of potentially thousands of parallel tasks, interconnected to one another with communication channels. Such applications are suited to be mapped onto manycore processors, which basically are sets of processors with local memory, interconnected with a network on chip (NoC).

Several tasks may be allocated to the same processor, as long as the total resource usage (memory, CPU, IO…) of all tasks allocated the processor does not exceed the available quantities. Communication between tasks in the same processor will use the internal memory, while routes on the NoC must be established for communication between tasks on different processors. The goal is then to reduce the global NoC usage by maximizing the in-processor communication.

Figure 1 illustrates the mapping of the tasks from a grid of 18x18 tasks onto 9 processors interconnected with a 2D torus network. Each processor has an assignment limit of 40 tasks. The different resulting colors correspond to the different processors the tasks are allocated to.

As shown on Figure 2, speedups on more than 30 could be experienced on instances of more than 500 tasks. We could find very good solutions for instances of more than 1000 tasks in less than 10 minutes, making the method suitable for late development processes, where final builds of applications are to be packaged for the embedded platform.

References:
A Heuristic Algorithm for Stochastic Partitioning of Process Networks

ABSTRACT: In this work, we study the problem of partitioning networks of processes under chance constraints. This problem arises in the field of compilation for multi-core processors. The theoretical equivalent for the case we consider is the Node Capacitated Graph Partitioning with uncertainty affecting the weights of the vertices. For solving this problem we propose an approximate algorithm which takes benefit of the available experimental data through a sample-based approach combined with a randomized greedy heuristic, originally developed for the deterministic version. Our experimental results illustrate the algorithm ability to efficiently obtain solutions of good quality within an acceptable execution time which are also robust to data variations.

The development of 100+ cores microprocessor architectures has triggered a renewed interest for the so-called dataflow programming models in which one expresses computation-intensive applications as networks of concurrent tasks interacting through (and only through) unidirectional FIFO channels. In [1], we present a heuristic algorithm dedicated to the resource-constrained graph partitioning problem which crops up when mapping networks of dataflow processes on a parallel architecture assuming the resource consumptions of the processes are uncertain.

Known, in the deterministic case of single dimensional weights, as the NP-hard problem of Node Capacitated Graph Partitioning, the assignment of the weighted vertices of a dataflow graph to a fixed set of partitions, has, to the best of our knowledge, received little attention from the stochastic programming community.

In order to respect as close as possible the real context of our application, a qualitative analysis of the sources of uncertainty, mainly the execution times, was performed. This preliminary analysis showed the inherent difficulty of obtaining an analytical description of the distributions of the execution times. Even if it is reasonable to assume that the probability distributions of execution times have a bounded support (no infinite loops), we have to cope with the fact that these distributions are intrinsically multimodal (due to the presence of data dependent control). Also, in the case of process networks, we cannot overlook the problem of dependencies between these random variables.

The approach we propose is justified by the theory of statistical hypothesis testing and takes into account the important role of experimental data. Additionally, for solving a chance constrained problem, no assumptions are being made about the joint distribution of the random variables, in particular with respect to the independence of these variables.

Our algorithm design methodology consists in leveraging an existing heuristic for the deterministic case without significant destructuring (i.e. at small cost in terms of software engineering) and with acceptable performance hit. Furthermore, this non-parametric method we introduce for solving our chance-constrained partitioning of process networks is applicable, in combination with other approximation algorithms (e.g. metaheuristics), to other optimization problems.

Figure 1 shows the statistically significant approximation model to an initial chance-constrained program, obtained with our robust binomial approach. $X_i$ are binaries variables and their sum follows a Binomial distribution, $k$ is a parameter determined in function of $NS$, $c$ and $\alpha$, and $L$ is a constant of large size, depending on the problem structure but generally easy to find.

For our problem, the objective $f$ is to minimize the communication inter partitions and the probabilistic constraints are on the capacities of the clusters. The heuristic we adapted for treating the stochastic graph partitioning is an already available greedy randomized affinity-based heuristic, easy to modify and quite efficient for the placement of the processes in the deterministic case. The task weights are random variables (memory footprint or computing core occupancy) for which we dispose of a relevant sample of $NS$ independent and identically distributed realizations.

By using the statistical hypothesis testing within a heuristic approach, we overcome the computational effort of taking into account the uncertainties of the weights of the vertices. Concerning the complexity, we remark a linear increase with a factor of $NS$ in comparison to the deterministic version. This approach can solve, with an acceptable solution quality, confidence level and computation time, problems representative in size of our application context. The overall solutions have a quality comparable to those of the heuristic for the deterministic case and moreover they are statistically guaranteed at a confidence level $1-\alpha$.

References:
A Low-overhead Dedicated Execution Support for Stream Applications on Shared Memory CMP

P. Dubrulle, S. Louise, R. Sirdey, V. David

ABSTRACT: The ever growing number of cores in Chip Multi-Processors (CMP) brings a renewed interest in stream programming to solve the programmability issues raised by massively parallel architectures. Stream programming languages are flourishing (StreamIt, Brook, ΣC, etc.). Nonetheless, their execution support have not yet received enough attention, in particular regarding the new generation of many-cores. In embedded software, a lightweight solution can be implemented as a specialized library, but a dedicated micro-kernel offers a more flexible solution. We propose to explore the latter way with a Logical Vector Time based execution model, for CMP architectures with on-chip memory.

Many-cores represent a challenge for programmers. Using efficiently the parallelism of hardware with hundreds of cores, using on-chip shared memory can be done with a language offering good abstraction and with an adapted compiler [1].

Stream programming languages (based on Kahn Process Networks) are suitable as they offer a guarantee for determinism, an abstraction of underlying hardware and good properties for efficient placing/routing tools [1].

Stream programming is a very good approach for signal and image processing, which are predominant in the embedded applications. For embedded many-cores, the execution support of stream applications is possible through an efficient execution model implemented as a micro-kernel, with a dynamic scheduling [2].

Figure 1: duration of system operations depending on the number of tasks in the executed application

An offline scheduler can infer a partial order of execution of the tasks in a stream application, as to guarantee determinism of data access. From this partial order of execution, it is possible to encode dependencies between all the activations of all tasks in a stream application by assigning them each a vector clock and increments to update them infinitely using addition and modulo operations. These vector clocks capture causality between computation events, comparing them tells if an event precedes another or if they are causally independent. For a given pair of tasks \(a\) and \(b\), checking current activation of \(a\) precedes the current activation of \(b\) using the vector clocks can be done by a scalar operation [2].

The proposed execution model relies on the vector clocks to update a data dependency counter at runtime. Initially, the counter’s value is known and tasks without dependencies are executed. When a task activation ends, the vector clocks are compared and dependencies are updated according to the comparison results.

In the micro-kernel, these operations are performed on a supervision core (many-cores tend to propose such additional cores for housekeeping, like the MPPA chip). This asymmetric approach takes advantage of the target parallelism to absorb a part of the scheduling overhead [2].

The micro-kernel is designed to run on the multi-core scale (up to 16 cores). To scale up to the many-core architectures, several instances of this multi-core micro-kernel run on partitions of the global set of cores (either logical partitions, or physical clusters in a hierarchical architecture) [2].

A prototype of this execution support was realized and evaluated on a x86 multi-core platform. The performance results show that the execution support’s overhead is not depending on the count of tasks in the application, which means the micro-kernel offers scalable dynamic scheduling as can be seen on figure 1. This evaluation demonstrates that the LVT micro-kernel is an efficient execution support for stream applications with filters implementing complex computations. On the other hand, applications with filters performing short operations require the intervention of compilation tools to merge some parallel tasks to reach an appropriate minimum execution time [3].

References:
Autonomic Pervasive Applications Driven by Abstract Specifications

Research topics: Internet of Things, Service-oriented & Autonomic Computing

O. Gunalp, L. Gurgen, V. Lestideau (LIG), P. Lalanda (LIG)

ABSTRACT: Pervasive application architectures present stringent requirements that make their development especially hard. In particular, they need to be flexible in order to cope with dynamism in different forms (e.g. dynamically changing service and data providers and consumers). Existing development approaches do not provide explicit support for managing this dynamism. In this paper we describe Rondo, a tool suite for designing pervasive applications. Rondo proposes pervasive application specification, which borrows concepts from service-oriented component assembly, model-driven engineering (MDE) and continuous deployment, resulting in a more flexible approach than traditional application definitions.

Pervasive computing aims at removing the barrier between users and computing systems by blending the computers into the users’ environment. This vision is becoming possible in the near future thanks to recent evolution in mobile, wireless and sensor technologies. However, the development of pervasive applications is a difficult challenge because the developer needs to manage contextual changes, device and application dynamism, as well as the business logic. As a consequence, current pervasive applications are generally insufficient in terms of software engineering: they are difficult to design, code, test and maintain; most existing solutions are proprietary, limited in terms of provided services and executed in a closed world.

For this, we propose a tool suite, Rondo [1], which enables design, development and execution of dynamic pervasive applications using runtime models and autonomic computing principles. We adopt a service oriented approach, which aims to promote loose coupling between components. The implementations of the services are decoupled from their specifications. As so, applications can be built upon loosely coupled service providers and consumers based on service level contracts. Late binding and sustainability becomes possible opening the way of dynamically adaptable software architectures.

Rondo is a tool suite for designing, deploying and executing pervasive applications. Rondo framework uses a model-driven approach and has three main goals: designing dynamic applications, specifying pervasive environment and enabling application adaptations for context-awareness. By this approach we aim to manage life cycle of pervasive applications, from development until runtime changes. Rondo provides a domain-specific language based on the notion of components to define the architecture of pervasive applications. Then at runtime, the application manager takes this description and configures the service-oriented execution environment in order to deploy and start the application, all taking into account current state of the environment, represented by several runtime models (Figure 1). And while the application is running, this manager continues to monitor and manage the created application.

We have implemented and validated utility of our approach, demonstrating it in a smart home application in the context of a European project, namely BUTLER. The application enables the media streams to follow the user in the home in different rooms equipped with audio/video devices so that the user can continue to watch and/or to listen the media while he/she moves through the house. Several sensors such as presence detectors, sonar sensors, touch sensors, etc. based on heterogeneous technologies (ZigBee, 6LoWPAN/CoAP) are used to localize the user.

References:
The coordination middleware developed in our team aims to deploy [3] and coordinate hardware and software components that are not initially planned to talk together. Building Automation (BA) is one of the fields where such middleware can increase its usefulness.

Indeed, BA Systems (BAS) encompass a wide variety of systems, e.g. HVAC, lighting, access control, intrusion alarm, fire detection. A building that integrates several BAS can be seen as a System of Systems (SoS) where the integration and interaction of the systems together with their own individual control aims at providing a higher level of energy efficiency, user comfort, etc. Unfortunately, most of the BAS coexist without any cooperation and interoperability, leading to conflicting actions. Moreover, the individual control of each (sub-)system does not lead to a global optimum for the whole targeted objective. In addition well established industrial protocols (e.g. BACnet, LonWorks, KNX, ...) need to cohabit with the new wireless technologies more and more present in the field.

This leads to a complex heterogeneous network with a variety of technologies which make the task of the people in charge of the administration of the system a nightmare. To overcome these difficulties, a middleware can provide an abstraction view of the underlying systems and provide some facilities for coordinating all the individual networks. At the level of the abstraction level, we have developed a dedicated framework named PUTUTU[1] (see Fig. 1) in order to ensure the separation of concerns that allows delegating to the appropriate people the different tasks required when BA is considered.

The implementation of the "driver" between the gateway and the devices is let to the responsibility of a person that is familiar with the communication protocol and the specific aspects of the sensors and actuators. Once done, this "driver" is directly usable by the PUTUTU framework which provides the abstraction level. At this level there is no difference in between a wired LON sensor and a wireless sensor operating in the 433Mhz range. Both of them are seen as resources contained in bags that can be interrogated through the coordination protocol[1]

In addition the generic blocks presented in the green box (Fig 1) take care of the treatments common to all the sensors actuators technologies: initialisation of the gateway, frequency sampling, management of the basic information (type, timestamps, journalization, ...). Then the application designer, who can be a third person can define the scenarios thanks to the rule based coordination language with no need to know deep details of the PUTUTU framework and the technology used by the sensors and actuators. Low power consumption and reliability are two important properties of wireless sensor networks. To improve these aspects, we go one step further and enforce the coordination protocol on top of the communication protocols imposed by the different wireless sensor networks. Thus, we move the callee side of this protocol from the gateway to the sensors/actuators in order to make them able to directly respond to this protocol [2], see Fig. 2. The high-level coordination protocol brings on the one hand the control from the application side the activities (sleep/awake) of the sensors and on the other hand the transactional processing of operations involving a group of sensors/actuators. This has a positive impact on the consumption and on the reliability.

References:
Reliability & Test

Memories

Wire Diagnosis
ABSTRACT: Error-correcting codes (ECC) offer an efficient way to improve the reliability and yield of memory subsystems. ECC codeword length is not the maximum allowed by a certain check-bit number since the number of data-bits is constrained by the width of the memory data interface. This work investigates the additional error correction opportunities offered by the absence of a perfect match between the numbers of data-bits and check-bits in some of the most commonly used ECCs. A method is proposed for the selection of multi-bit errors which can become correctable with a minimal impact on decoder latency. Reliability improvements are evaluated for memories in which all errors affecting the same number of bits in a codeword are equally probable.

Error-correcting codes (ECC) provide an effective way to achieve the required level of transient fault tolerance in storage and memory subsystems since they can be applied at system or component levels with relatively limited design overhead. Since the implementation of an ECC requires a certain amount of storage overhead, any approach able to boost the fault masking capacity is helpful. An approach with lower performance overhead is to extend the error correction capability of an ECC without increasing the check-bit number per codeword. This is possible due to the fact that usually the number of data-bits needs to be a power of 2 or a multiple of a power of 2. For example, an (16, 8) ECC allows the correction of all single-bit and double-bit errors, which means 136 errors, while, in principle, up to 256 distinct errors could be distinguished with the available information redundancy.

Most of these ECC extensions are devoted to the correction of burst errors which affect contiguous codeword bits. This choice is not justified when the burst errors are not necessarily the most probable multi-bit errors. Examples include CMOS memories protected against multi-bit upsets with the help of bit interleaving or non-volatile memories where the corruption of information is not necessarily induced by ionizing particles such as the magnetic RAMs (MRAM).

Here, we propose the concept of maximized ECCs obtained by a better utilization of the information redundancy available in linear block ECCs. Starting from an N-bit ECC, the goal is to get a maximum number of correctable (N+1)-bit errors. Selection criteria are defined in order to reduce the impact on the latency of the error correcting logic. For example, among all correctable (N+1)-bit errors which generate the same syndrome, the errors which affect a maximum number of check-bits are privileged.

The impact on mean-time-to-failure (MTTF) of a memory was evaluated for the particular cases of ECCs that enable the correction of errors affecting a maximum of one bit or two bits. Especially in the case when two-bit errors can be corrected, significant MTTF increase can be obtained. In this paper, only reliability improvements are quantified despite the fact that such solutions could also be used to improve memory manufacturing yield.

Consider a memory unit protected by an N-bit ECC. Memory failures occur if codewords with more than N corrupted bits are accessed. It is assumed that:

- The (N+1)-bit errors are independent and identically distributed with a given rate \( \lambda \).
- The rate of errors affecting more than N+1 bits is much smaller than \( \lambda \).

In these conditions, the probability \( P \) that a memory unit operates without failure during a time interval \( \tau \) can be expressed as:

\[
P = e^{-\lambda \tau}
\]

where \( 1/\lambda \) gives the memory MTTF under the specified assumptions.

If a certain fraction \( x \) (0 ≤ x ≤ 1) of the (N+1)-bit errors can be masked by a certain mechanism, the (N+1)-bit error rate responsible for memory failures will become (1-x)\( \lambda \). And the memory MTTF improvement is given by the following expression:

\[
\frac{\Delta MTTF}{MTTF} = \frac{x}{1-x}
\]

If 20% of the (N+1)-bit errors become correctable, then the MTTF can be improved with 25%. This means that an MTTF of 4 years can be extended with one additional year. We generate H-matrices of the maximized double-bit error correcting codes (DEC) codes from scratch with the help of a SAT-solver. If only data-bits need to be provided by a maximized DEC decoder, the hardware overhead can be reduced by selecting those triplets that involve a maximum number of check-bit positions. Besides the constraints specific to a DEC code, two additional goals were imposed:

- Find an H-matrix that can provide a maximum number of triple-bit errors that can be corrected,
- Maximize the number of correctable triple-bit errors that involve only check-bit positions.

Once the H-matrix is found, the set of correctable triple-bit errors is constructed by selecting first those triple-bit errors which involve only check-bit positions. The next privileged triple-bit errors affect two check-bit positions. The remaining triple-bit errors are selected among those which involve one or zero check-bit positions. All triple-bit errors that affect contiguous bit positions can be made correctable without affecting the total number of correctable triple-bit errors.

Table I reports the ratios of triple-bit errors masked with the obtained maximized DEC codes with respect to the total numbers of triple-bit errors. The achieved MTTF improvements were between 27% and 100%. The maximized DEC decoders were synthesised with a 45nm standard cell library and their latency overheads were between 0% and 20%.

<table>
<thead>
<tr>
<th>DEC code</th>
<th>Number of data-bits</th>
<th>Number of check-bits</th>
<th>Achieved number of masked triple-bit errors and ratio of triple-bit errors</th>
<th>Achieved MTTF improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>(16,8)</td>
<td>8</td>
<td>8</td>
<td>118 21%</td>
<td>27%</td>
</tr>
<tr>
<td>(22,12)</td>
<td>12</td>
<td>10</td>
<td>770 50%</td>
<td>100%</td>
</tr>
<tr>
<td>(26,16)</td>
<td>16</td>
<td>10</td>
<td>672 26%</td>
<td>35%</td>
</tr>
<tr>
<td>(36,24)</td>
<td>24</td>
<td>12</td>
<td>3250 46%</td>
<td>84%</td>
</tr>
<tr>
<td>(44,32)</td>
<td>32</td>
<td>12</td>
<td>3100 23%</td>
<td>31%</td>
</tr>
</tbody>
</table>

Table I - MTTF of Maximized DEC Codes
A Distributed Diagnosis Strategy using Bayesian Network for Complex Wiring Networks

Research topics: wiring network, reflectometry, Bayesian Networks, uncertainty

ABSTRACT: In this paper, a distributed diagnosis strategy using reflectometry is proposed. It consists in making reflectometry measurements at different spots of a highly complex wiring network. The proposed approach targets sensors number optimization using Bayesian Networks. It consists in three steps: (1) sensors implementation in a deterministic case, (2) diagnosis modeling using Bayesian Networks, (3) sensors number optimization. Here, the main objective is to find a compromise between the sensors number (system cost) and diagnosis measure uncertainty (diagnosis quality).

Reflectometry is a powerful technique for electrical faults detection, localization and characterization. In branched wiring networks, a distributed diagnosis strategy is required to guarantee a good diagnosis quality. The main idea is to implement several sensors at different locations on the network. However, this multi-sensor architecture has imposed serious challenges on signal processing, sensors number and location optimization, resource allocation, etc. This paper focuses on sensors number optimization depending on a predefined target confidence level. The main objective is to find a good compromise between diagnosis quality and system cost. The proposed approach consists in several steps which are:

**Deterministic case implementation:**
One sensor is implemented at each end of the transmission line, which maximizes diagnosis coverage. Here, a high diagnosis quality is obtained (i.e. 100% confidence level), but the system cost is also very important.

**Diagnosis modeling using Bayesian Networks (BN):**
In order to reduce BN implementation complexity, a local BN is modeled for each sensor as depicted by Fig.1.

Then local BNs are integrated into a global BN in order to locate faults on the whole network [1].

**Sensors number optimization:**
In order to reduce sensors number, the network is divided into several generic networks of Y or star topology. Each sub-

network is diagnosed by only one sensor as depicted by Fig.2.

A fault is simulated on branch B3 of the network. In this case, a low system cost is achieved (from 6 sensors down to 3). However, the obtained diagnosis quality is really bad (confidence level is equal to 33%). In order to overcome this problem without adding any sensors in the network, communication between neighboring sensors is introduced to exchange information about the detected fault. Fig.3 shows the global Bayesian network model in the optimized case. Here, diagnosis quality and confidence level are both satisfactory (equal to 100%).

As future works, the cable life profile will be introduced for better sensors optimization.

References:
Soft Faults Diagnosis in Wire Networks Using Time Reversal Reflectometry

Research topics: FDTD, Reflectometry, Time Reversal.

L. El-Sahmarany, F. Auzanneau, L. Berry\textsuperscript{1} and P. Bonnet\textsuperscript{1} (\textsuperscript{1}Université Blaise Pascal)

**ABSTRACT:** The invariance of the wave equation under time-reversal (TR) in lossless transmission line is exploited for detection and localization of soft faults in a wire network. A TR-based signal processing is exposed and evaluated on numerical examples. To test the efficiency of this method, the TR algorithm has been developed and simulated using FDTD (Finite Difference Time Domain Method). It allows us to better diagnose soft fault in the wire thanks to the time reversal method.

Reflectometry methods are commonly used for testing transmission lines. Hard faults (open and short) are observable by standard reflectometry, but soft faults (damaged insulation, etc...) are generally not (as illustrated in Fig. 1). This study presents a new signal processing based on time reversal for the detection and the localization of soft faults.

In the reverse temporal space, a convolution product is then calculated $V_{\text{bis,rd}}*V_{\text{in}}$. This operation reaches a maximal value for the component of the line which has a time delay equal to the time delay of the fault. Fig. 3 represents the convolution result which allows focusing the energy on each fault without a priori knowledge on their locations. A modification of a per-unit-length parameter, respectively 38\% of $C$ (capacitance) and 11\% of $L$ (inductance) was supposed to simulate each fault.

Time reversal was first introduced in acoustics by M. Fink. This technique efficiently focuses energy on a target taking benefit from the invariance property of the propagation equation with respect to time.

Adopting a similar approach, any local impedance discontinuity in a transmission line, created by a soft fault, behaves like a secondary source generating a transmitted wave and a reflected wave. Thus, a TR process can be applied to locate these modifications relative to a healthy reference line.

As illustrated in Fig. 2, the proposed signal processing requires three steps. A voltage pulse is injected in the cable without fault and propagates according to the wave propagation equations. The simulation method gives the spatial voltage distribution along the transmission line: $V_{\text{in}}$. In a second step, the same pulse is injected in the cable with fault and the reflected signal $V_{\text{rd}}$ is recorded. In a third step, the recorded signal $V_{\text{rd}}$ is time-reversed and re-injected into the cable without fault. The spatial voltage distribution $V_{\text{bis,rd}}$, along the line is provided by the simulation code.

This method improves the efficiency of detecting and locating “soft faults” in a transmission line or in a wire network.

**Reference:**

Time Reversal Reflectometry for Cable Ageing Characterization

Research topics: Reflectometry, Time Reversal, Cable ageing.

L. El-Sahmarany, F. Auzanneau and P. Bonnet (Université Blaise Pascal, Clermont-Ferrand)

ABSTRACT: We investigate the effects of ageing (i.e. slow homogeneous degradation) on electrical cable characteristics by the use of a new method based on time reversal. In case of a global cable ageing, the commonly used methods such as reflectometry provide non-relevant or inaccurate information. Through theoretical study and numerical simulations, the benefits of this new method called Time Reversal Reflectometry (TRR) are presented. TRR is experimentally shown to be effective for the detection and quantification of cable ageing.

Ageing is described as a slow structural modification which gradually decreases the efficiency of an object, information or organism to provide its functions. Therefore, this paper overcomes reflectometry’s limitations by proposing a new approach based on time reversal applied to reflectometry’s fundamental principle. It focuses on the detection and estimation of electrical cable ageing.

This new method is based on the principles of time reversal and standard reflectometry methods [1]. Instead of using a predefined signal (Gaussian pulse) like standard reflectometry, it uses an adapted signal that allows characterizing more precisely the cable’s electrical parameters (RLCG) modifications due to ageing. The adapted signal will be insensitive to dispersion which distorts signals and decreases the ability of cable ageing detection and estimation. The detection of cable ageing using time reversal is summarized by the following process:

1) Inject a (symmetrical) pulse signal into a healthy cable.
2) If needed, truncate and shift in time, then normalize the reflected signal.
3) Apply time reversal, and then save as adapted signal.
4) Inject the adapted signal into the aged cable.
5) Process the reflected signal, calculate the Skewness Coefficient noted SC, and estimate the cable ageing. SC is calculated by quantifying the signal’s distortion on the left “a” or the right side “b” of its maximum as presented on Fig.1. Then, SC = b/a. A value of SC close to 1 means the cable under test is healthy. Otherwise, if the reflected signal is asymmetrical or SC value is far from 1, this means the cable is aged and the value of SC enables to quantify the ageing.
6) Loop steps 4 and 5 when needed.

Step 5 for the healthy cable (Capacitance is C0) and for three simulated aged cables (Capacitance values of aged cables are 0.2, 0.6 and 1.2 times C0) were calculated. Table I presents the values of SC and shows the effect of ageing. When the cable is healthy, SC is equal to 1 and when it is aged (0.2 * C0) SC is down to 0.5127.

<table>
<thead>
<tr>
<th>Simulations</th>
<th>1.2*C0</th>
<th>C0</th>
<th>0.6*C0</th>
<th>0.2*C0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>1.09</td>
<td>1</td>
<td>0.7943</td>
<td>0.5127</td>
</tr>
</tbody>
</table>

Thermal ageing experiment was performed on a 100 m long coaxial cable [2]. Table 2 shows the effect of ageing via the variation of the skewness coefficient. It was noted that the increase of SC with time led the reflected signal to lose its symmetry (as illustrated in Fig.2).

In order to investigate this method a comparison by changing the per unit length capacitance value was performed, the simulation is done by using a RLCG frequency model of a cable using MATLAB®. The reflected signal from the proposed method presents a simple and more accurate technique to estimate cable ageing. It can help monitor the health of the cables and the safety of an entire electrical system.

References:
New Advances in Monitoring the Ageing Of Electric Cables in Nuclear Power Plants

Research topics: Cable Ageing, Reflectometry, Signal Processing, Nuclear Plant

M. Franchet, N. Ravot, N. Gregis, J. Cohen, O. Picon (Université Paris-Est, ESYCOM)

ABSTRACT: Monitoring the ageing of electrical cables used in nuclear power plants is a crucial issue for nuclear industrials, whose objective is to extend the lifetime of their plant while ensuring its security. For cost and efficiency reasons, reflectometry, which is a non-destructive method, is well-adapted to this problem. Unfortunately, it may be not sensitive enough to small changes of the cable. To overcome this difficulty this article proposes to use time-frequency tools (the Wigner Ville transform and a normalized time-frequency cross-correlation function) in addition to time domain reflectometry. This method has been applied on two RG-59B coaxial cables (a new one and an old one) commonly used in nuclear power plants.

Studying cable ageing involves being able to detect minor modifications. This is the same challenge as detecting soft faults (incipient defects), which translate into reflected signals of very low amplitudes. Indeed only slight changes due to ageing will affect the reflectograms obtained by Time Domain Reflectometry (TDR), Fig 1. To detect them, TDR can’t be used alone, another tool is needed. In this article, we propose to apply a time-frequency transform, called the Wigner Ville transform (WVT), on TDR results and compute a normalized time-frequency cross-correlation function (TFC), Fig 2. This is part of a method called Joint Time-Frequency Domain Reflectometry (JTFDR) which has shown promising results for soft faults. However the WVT is a quadratic transform. Then unwanted cross-terms can affect the results. To overcome this problem the Pseudo Wigner Ville transform (PWVT) can be used instead of the WVT [1]. JTFDR has already been used to study local ageing of cables used in NPPs. This is comparable to detecting soft faults. Nevertheless the ageing process may affect the entire length of the cable. So the efficiency of such a method has to be tested for global ageing too and a tool to decide if a cable has to be replaced or not must be defined.

Table 1 gives the results obtained, for two kinds of injection, when the TFC is computed with the WVT and the PWVT. Using a thinner injected signal makes it easier to determine if a cable has been used or not.

Table 1: Correlation coefficients obtained depending on the kind of injection

<table>
<thead>
<tr>
<th>Width of the injected pulse</th>
<th>Correlation coefficient obtained with the WVT</th>
<th>Correlation coefficient obtained with the PWVT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ns</td>
<td>0.57</td>
<td>0.65</td>
</tr>
<tr>
<td>5ns</td>
<td>0.96</td>
<td>0.94</td>
</tr>
</tbody>
</table>

This method performs well on the studied example. However this study has to be deepened in order to see if it is able to discriminate different level of ageing. This is the object of current and future work made within the framework of Advance, a FP7 European program.

References:
On a Useful Tool to Localize Jacks in Wiring Network

Research topics: wired network diagnosis, soft fault detection

M. Franchet, N. Ravot, O. Picon (Université Paris-Est, ESYCOM)

ABSTRACT: To efficiently monitor and maintain wired networks, their topology has to be known. Most of the time a wiring network is made of several cables linked to each other with connectors. So knowing where they are is valuable information for going back to the topology. Then the damaged portions of the network can be localized relatively to the jacks, which will facilitate and accelerate maintenance. New data processing techniques based on a time frequency transform are shown to improve the detection capacity of both soft defects and connectors in wires.

Electrical cables are everywhere in many fields where the transfer of energy and information is necessary to guarantee the performance of a system. One day or another, a cable network will show signs of weakness or ageing involving the appearance of defects. These anomalies can be at the origin of dysfunctions and imply serious consequences for the system or the environment. This is why diagnosis methods for wired networks have been thoroughly studied in the past few years.

Reflectometry based methods have proven to be the best suited, as they provide detection and localization information, while requiring only one connection to the network. But, Time Domain Reflectometry (TDR) or Frequency Domain Reflectometry (FDR) methods are well suited for hard defects (i.e. defect that prevent any signal from going further away). But soft defects, such as localized damage to the insulation or shielding of a wire, are much more difficult to diagnose. This kind of defect account for 30% to 50% of all detected wiring faults, and is the premises of future hard defects.

A new method, called JTFDR (Joint Time Frequency Domain Reflectometry) takes benefit of the advantages of both TDR and FDR while avoiding their limitations by the use of innovative signal processing. It is based on the use of the Wigner Ville transform (WVT) coupled to a normalized Time Frequency Cross-correlation function (TFC) applied to TDR measurements, which greatly enhances the connector’s and soft defects signatures.

Fig. 1 shows that standard TDR measurement cannot efficiently detect and locate a connector or a soft defect in a line, as their peaks have very weak amplitudes [1]. The WVT has previously shown great ability for time-frequency localization of chirp-like signals. For this reason, the Pseudo WVT is combined with a normalized Time Frequency Cross-correlation function (TFC), defined below.

$$C_{sr}(t) = \frac{2\pi}{E_r(t) \cdot E_s} \int_{t_T-1}^{t_T+1} \int_{-\infty}^{+\infty} PW_r(t', \omega) \cdot PW_s(t'-t, \omega) d\omega dt'$$

In this formula, Es(t) and Er are normalization factors. The first normalization term provides local amplification of the weak signals isolated by the WVT. As a result, the weak peaks of the connector and the soft defect are enhanced to a level similar to the end of line reflection (an open circuit), making them much easier to detect and localize [2].

Fig. 2 shows the enhancement of the connector’s and the soft defect’s signatures using TFC (from [2]).

So knowing where these jacks are is valuable information for going back to the topology. Once this information is known, the damaged portions of the network can be localized relatively to the connectors; this greatly facilitates and accelerates maintenance.

Besides, in a wiring network, the jacks themselves can be sources of damage. So it is important to be able to monitor their condition. Knowing where they are is the first step to study their health and anticipate their ageing.

References:


PhD Degrees Awarded

Maud Franchet
Céline Azar
Sébastien Courroux
Olivier Bichler
Fabien Gavant
Mykhailo Zarudniev
Maud Franchet
University: Université Paris-Est
Reflectometry applied to soft fault detection in bundles of wires

The research works presented in this thesis is about the topic of detecting soft faults (incipient faults) in specific wiring structures: multiconductor transmission lines (MTL), also known as bundles of wires. The reflectometry methods, often used for the diagnosis of wiring networks, aren’t for now efficient enough to detect such defects. Besides, they have been designed for single lines only, where electromagnetic coupling between conductors (crosstalk) is mostly irrelevant. However such phenomenon can provide more information about the state of the cable. Using this information could enable us to detect soft faults more easily. In this work we propose a new reflectometry method, which takes advantage of crosstalk signals in order to detect incipient faults. Such a tool has also the advantage of being well-adapted to bundles of cables.

Thanks to the preliminary study of the impact of soft faults on the characteristic parameters of a multiconductor transmission lines and on crosstalk signals, a method called « Cluster Time Frequency Domain Reflectometry », has been proposed. It is a three step process. First temporal reflectometry measurements are made at the beginning of the line under test. All the available signals, even crosstalk ones, are recorded. A time-frequency process is then applied on them, in order to amplify the presence of defects. Finally, a clustering algorithm, that has been specifically developed for wiring diagnosis, is used to benefit from the whole available information.

Céline Azar
University: Université de Bretagne Sud
On the design of a distributed adaptive manycore architecture for embedded systems

Chip design challenges emerged lately at many levels: the increase of the number of cores at the hardware stage, the complexity of the parallel programming models at the software level, and the dynamic requirements of current applications. Facing this evolution, this PhD thesis aims at designing distributed adaptive manycore architecture, named CEDAR (Configurable Embedded Distributed ARchitecture), which main assets are scalability, flexibility and simplicity. The CEDAR platform is an array of homogeneous, small footprint, RISC processors, each connected to its four nearest neighbors. No global control exists, yet it is distributed among the cores. Two versions are designed for the platform, along with a user-familiar programming model. A software version, CEDAR-S, is the basic implementation where adjacent cores are connected to each other via shared buffers. A co-processor called DMC (Direct Management of Communications) is added in the CEDAR-H version, to optimize the routing protocol. The DMCs are interconnected in a mesh fashion.

Two novel concepts are proposed to enhance the adaptiveness of CEDAR. First, a distributed dynamic routing strategy, based on a bio-inspired algorithm, handles routing in a non-supervised fashion, and is independent of the physical placement of communicating tasks. The second concept presents dynamic distributed task migration in response to several system and application requirements. Results show that CEDAR scores high performances with its optimized routing strategy, compared to state-of-art networks. The migration cost is evaluated and adequate protocols are presented. CEDAR is shown to be a promising design concept for future manycores.
Sébastien Courroux
University: Université de Bourgogne

Wavelet-based algorithms for embedded image processing and integration into a smart vision system.

Data at the output of a CMOS image sensor are processed through a set of operations, either for purposes of rendering or image analysis. Increasing resolution of the sensors and reducing of the size of the pixels make operations to be applied even more complex and require a large storage capacity. It is increasingly difficult to reconcile these different constraints in a low-cost embedded sensor, consisting of an analog part and a digital circuit having a single processor, low storage capacity and low operating frequency. New methods are then investigated. One of them proposes to use alternative data representation. The wavelet representation decomposes an image into frequency bands, orientation and scale, simplifying the future operations of the processing chain.

In a first step, the thesis proposes to study the interest of the wavelet representation for image processing in embedded real time context. For this, a state of the art of the algorithm methods is established and allows defining two algorithmic chains: reconstruction of CFA images and facial recognition. The quality of the process is demonstrated for these two processing.

In a second step, a wavelet-oriented vision system is proposed consisting of an embedded processor and a module dedicated to the wavelet transform. The wavelet transform module adopts a so-called 'semi-folded' structure and performs effectively the wavelet decomposition at several scales using only a few lines of internal memory. This vision system is used to speed up processing and increase application flexibility and effectiveness of low-cost sensors.

Olivier Bichler
University: Université Paris-Sud

Adaptive Computing Architectures Based on Nano-fabricated Components

In this thesis, we study the potential applications of emerging memory nano-devices in computing architecture. We show that neuro-inspired architectural paradigms could provide the efficiency and adaptability required for complex image/audio processing and classification applications with a much lower cost in terms of power consumption and silicon area than current solutions. This work is focusing on memristive nano-devices, such as: Phase-Change Memory (PCM), Conductive-Bridging RAM (CBRAM), resistive RAM (RRAM).... We show that these devices are particularly suitable for the implementation of natural unsupervised learning algorithms like Spike-Timing-Dependent Plasticity (STDP), requiring very little control circuitry. The integration of memristive devices in crossbar array could provide the huge density required by this type of architecture (several thousand synapses per neuron), which is impossible to match with a CMOS-only implementation.

In this work, we propose synaptic models for memristive devices and simulation methodologies for architectural design exploiting them. Novel neuro-inspired architectures are introduced and simulated for natural data processing. They exploit the synaptic characteristics of memristives nano-devices, along with the latest progresses in neurosciences.

Finally, we propose hardware implementations for several device types. We assess their scalability and power efficiency potential, and their robustness to variability and faults, which are unavoidable at the nanometric scale of these devices.
Fabien Gavant
University: Université de Grenoble

Architectures for Image Sensors Stabilization based on Visual Perception and on the Physiology of Hand Tremor; a Contribution

With the integration of cameras in mobile devices, their democratization and the reduction of the imager size, the optical system dimensions and the pixels miniaturization, the pictures become more and more subject to motion blur due to the hand tremor. In addition, the requirements in terms of image quality become higher and higher. Hence, in order to reduce this blur, several image stabilization systems have been developed. Nevertheless, they cannot guarantee the sharpness quality of resulting images and in some cases, they show integration difficulties.

In order to overcome these limitations, the research work presented in this thesis proposes, first of all, a physiological tremor model that aims at simulating realistic camera shake and secondly, presents a study on visual perception of blur. This study enables the development of a quality metric. Finally, stabilization algorithms and architectures exploiting these new tools are presented. These new architectures reduce the number of external components and ensure sharp stabilized images.

Mykhailo Zarudniev
University: Université de Lyon – Ecole Centrale de Lyon

Frequency Synthesis using Spin Torque Oscillator Coupling

Current trends in telecommunication are leading to multiple standards systems. The conventional solution consists in using one local oscillator for each standard. The spin torque oscillator (STO) is a new device that appears as a potential candidate for the LC-tank oscillator replacement, due to its wide frequency accordability and its small volume. However, it exhibits poor power and phase noise performance.

In this work, we propose to reach the technical specification of the radiofrequency applications by coupling a large number of spin torque oscillators. An original oscillator network model that describes qualitative properties of the oscillator synchronization is introduced. Next, the control law architecture for an oscillator set is established in order to achieve the technical specifications. Finally, we propose two original frequency domain design methods allowing the resolution of our frequency synthesis problem. The first design method allows considering explicitly a performance criterion corresponding to a desired frequency constraint. The method allows obtaining a suitable sub-system interconnection matrix that fits the frequency specification constraint. The second design method allows to find an interconnection matrix and to take into account simultaneously several frequency specification constraints. The interconnection matrix obtained with the proposed method solves the problem of frequency synthesis by coupling of spin torque oscillators.
Greetings

Editorial Committee

Marc Belleville
Christian Gamrat
Fabrice Auzanneau
Ernesto Perea
Hélène Vatouyas
Jean-Baptiste David

Readers

Thierry Collette
Eric Mercier

Graphic Art

Valérie Lassablière