Annual Research Report 2014

Architecture and IC Design, Embedded Software
Leti is an institute of CEA, a French research-and-technology organization with activities in energy, IT, healthcare, defense and security. By creating innovation and transferring it to industry, Leti is the bridge between basic research and production of micro- and nanotechnologies that improve the lives of people around the world. Leti partners with large industrials, SMEs and startups to tailor advanced solutions that strengthen their competitive positions. It has launched more than 50 startups. Its 8,500m² of cleanroom space feature 200mm and 300mm wafer processing of micro and nano solutions, for applications ranging from space to smart devices. Leti’s staff of more than 1,800 includes 200 assignees from partner companies. Leti is based in Grenoble, France, and has offices in Silicon Valley, Calif., and Tokyo. Visit www.leti.fr for more information.

List, an institute of CEA, 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 700 researchers, engineers and technicians, the CEA-LIST performs innovative research in partnership with major industrial players in the fields of ICT, Energy, Transport, Security & Defense, Medical and Industrial Process. List is based in the Paris-Saclay campus, France. Visit www-list.cea.fr for more information.

Design, Architectures & Embedded Software research activity is shared between Leti and List through a dedicated division. More than 245 people are focusing on RF, digital and SoC, imaging circuits, design environment and embedded software. These researchers perform work for both internal clients and outside customers, ranging from startups and SMEs to large international companies.
Contents

Editorial 5

Key Figures 7

Scientific Activities 9

Energy, Sensors & Diagnosis 11

Wireless Communications & Cyber-Physical Systems 25

Energy, Delay & Thermal Optimized Digital Design 39

Multi/Many-Cores Architectures & Software 53

Dynamic & Real-time Software 61

Emerging Technologies & Paradigms 71

PhD degrees awarded 81
Our field of Integrated Circuits and Embedded Systems design is and will be, a key to the on-going digital revolution, in particular to ensure a real-time coupling between the physical mode and the virtual world. Thus, we have to develop always more integrated, autonomous and smart devices which interact with the physical world, and simultaneously, solutions that collect, store data, and process it in real time in order to perform this essential coupling. This will enable this revolution and the move from "Big Data" to "Fast Data".

A new generation of more efficient computing solutions, especially in energy, is required; Neurocomputing, and adequate computing are some examples of it. In addition to a large autonomy, these systems must provide a high level of quality of service, particularly in terms of reliability, dependability and security, and thus, we are aiming towards "Real Time Cyber-Physical Systems".

Applications asking for these solutions are in the fields of transport, Internet of Things, security, medical, manufacturing (Industry 4.0), etc…

These issues are driving our current research activities. Some results will be integrated in our L-IoT project aiming to provide hardware and software demonstrators for autonomous smart devices, others in our Green&Secure Computing Module demonstrator dedicated to servers.

We hope you will appreciate reading this report that gives you an overview of our latest researches.

Thierry Collette
Key figures

2 locations:
MINATEC Campus (Grenoble)
PARIS-SACLAy Campus (Palaiseau)

190 Permanent researchers,
55 PhDs and Post-docs

Full suite of IC CAD tools,
Hardware Emulators,
& Test equipments,

35M€ budget
85% funding from contracts

48 granted patents
38 papers, journals & books
180 conferences & workshops
Publications
218 publications in 2014, including journals and Top conferences like ISSCC, ESSCIRC, IMS, ISCAS, DAC, DATE, SPIE, IPDPS, HIPEAC, CDC, ECC, and ESWeek

Prize and Awards
Best Presentation Award IECON 2014 - N.M. Nguyen et al.
Best Poster Award, DATE 2014 - F. Beneventi et al.
Best Paper Award, MCSoC-14 - S. Louise et al.
Microwave Prize, EuMC 2014 - C. Jany et al.
Silver Leaf Award, IEEE PRIME 2014 - R. Grezaud
Bronze Leaf Award, IEEE PRIME 2014 - A. Ratiu
ENIAC 2014 Innovation Award, EnLight Project,

Experts
45 CEA experts: 4 Research Directors, 2 International Experts
10 Researchers with habilitation qualification (to independently supervise doctoral candidates)
2 IEEE Senior Members

Scientific Committees
19 members of Technical Programs and Steering Committees in major conferences: ISSCC, ESSCIRC, DAC, DATE, ASP-DAC, ISPLED, ICCAD, IJCNN

Conferences and Workshops organizations
22nd International Conference on Real-Time Networks and Systems Conference, Versailles 2014
10th Conference on PhD Research In Microelectronics and Electronics, Grenoble 2014

International Collaborations
Collaborations with more than 20 universities and institutes worldwide
Caltech, University of Berkeley, University of Columbia, Carnegie Mellon University, EPFL, ETHZ, CSEM, UCL, UNIBO, Polito Torino, KIT, Chalmers University, Tongji, ...
Energy, Sensors & Diagnosis

Energy conversion
Imagers
Wire diagnosis Sensors & Signal Processing
An Autonomous Piezoelectric Energy Harvesting IC based on a Synchronous Multi-Shot Technique

P. Gasnier, J. Willemin, S. Boisseau, C. Condemine, S. Robinet

ABSTRACT: this work presents a fully autonomous Integrated Circuit (IC) that efficiently converts the AC output power of mechanical Energy Harvesters into a low-voltage DC source for supplying autonomous sensor nodes. The IC operates with high-voltage harvesters (from 3 to 200 V) using low-voltage AMS 0.35µm CMOS technology. It performs an efficient energy extraction thanks to an off-chip inductive element and a precise control strategy. It autonomously harvests mechanical energy and converts it into electrical energy from 10 µW to 500 µW at 3V-DC. Furthermore, the system self-starts and does not need any battery to operate.

In dusty or dark environments submitted to shocks, stresses or vibrations, where solar energy is unreliable or non-existent, mechanical energy harvesting is a relevant candidate to power Wireless Sensor Nodes (WSN). The piezoelectric principle is relevant to convert mechanical energy into electrical energy: in addition to its low cost and its availability, it offers a high power density at micro and macro scales and low mechanical frequencies. The output power of piezoelectric harvesters is generally in the 10 µW to 1 mW range and their high AC output voltages are not compatible with standard integrated circuits requirements. As a consequence, a power management circuit (PMC) is essential to turn the piezoelectric raw output into a viable supply source for WSNs. Synchronous Electric Charge Extraction (SECE) is a type of technique which enhances the energy extraction compared to a full-bridge rectifier and a smoothing capacitor. SECE consists in entirely discharging the piezoelectric harvester in a coil when its voltage reaches a maximum.

Our system is based on two energy paths to implement a self-starting battery-free operation mode. A Flyback is used to perform SECE (by precisely controlling Kp and Ks when the piezoelectric voltage reaches a maximum). Moreover, the PMC described herein implements a new SECE technique called MS-SECE (“Multi-Shot” SECE). Instead of a single transfer, the piezoelectric harvester is discharged in N successive energy transfers. This novel technique enables to reduce the volume of the power circuit and/or to improve the efficiency of the energy transfer compared to SECE, by reducing conduction losses.

The IC has been fabricated in the 3.3V AMS 0.35µm CMOS technology. The IC, surrounded by a few off-chip components, self-starts and operates with capacitors as storage elements. Thanks to the off-chip Flyback topology, it is not limited in terms of tolerated piezoelectric voltages. Fig. 2 shows in (1) the cold-start of Casic (10 µF) through Kd, in (2) the charge of Cwsn by MS-SECE (during which VDDasic decreases) and in (3) the operation of a WSN powered when Cwsn (100 µF) is filled. The WSN is based on a Texas Instrument CC430 platform consuming 100 to 200 µJ per operation, including a temperature measurement and a 868 MHz RF transmission.

It is shown that MS-SECE increases the efficiency of the energy transfer by 15% and up to 25%. The chip also enables to use small inductive components to discharge a considerable quantity of energy without reaching the saturation induction.

This IC, designed in a low voltage and low cost technology, consumes 1 µW @ 5 Hz. Along with the MS-SECE technique, it paves the way towards size reduction of power converters for piezoelectric harvesters while using efficient harvesting techniques: its overall volume is less than 2 cm³ (Fig. 3).

Related Publications:
Heterogeneous Integrated Voltage Level Shifter for High-Voltage, High-Frequency and High-Temperature Applications

Research topics: Power Electronics, Heterogeneous Integration, Wide Bandgap Devices

R. Grezaud, F. Ayel, N. Rouger (G2ELAB), J.-C. Crebier (G2ELAB)

ABSTRACT: This work presents a level shifter structure integrated to a SiC JFET-based power module and SOI gate drivers. High-side and low-side digital input signals are isolated from the power application by the same low-side 2-channel coreless isolator. One isolated signal is translated to the high-side potential reference by an additional SiC device driven by the low-side SOI circuit. The proposed voltage level translator architecture has been validated with SiC vertical JFETs and low-voltage MOSFETs. At 240V power supply voltage and 100 kHz switching frequency the level shifter propagation delay is only 10 ns for 1.85 mA current consumption.

Wide bandgap power devices like SiC JFETs or GaN HEMTs are very promising to address at the same time high-voltage, high-frequency and high-temperature applications. To convert power in such conditions power switches but also gate drivers and isolators have to operate properly. One of the most challenging devices is the digital isolator. It has to deal with high-voltage, high common-mode transient immunity, short propagation delay and high-temperature. To reduce price, volume and improve power system reliability we propose the inverter leg architecture depicted on Fig. 1. This one is based on an integrated high-speed level shifter and a single 2-channel isolator. By using a single isolator chip instead of two, the channel-to-channel matching is better and constraints on dV/dt immunity are loosened, but the high-side isolated input signal is now referenced to the low-side. Thus, an effective high-voltage level shifter is required to translate signal IN1" from the power ground GND to the floating high-side potential reference VSw.

\[ I_{R2} = I_{R1} = \frac{V_{IN} - V_{IN1\prime} - V_{th,J3}}{R_1} \]

The proposed level-shifter architecture has been validated with discrete 1200V normally-on SiC JFETs and 50V MOSFETs. Experimental waveforms are shown on Fig. 2. At 240V power supply voltage, 100 kHz switching frequency and 250 ns pulse width, the level shifter propagation delay is only 10 ns while it consumes 1.85 mA on the high-side driver power supply.
Noise optimization for low light CMOS image sensors

Research topics: CMOS imagers

A. Boukhayma, A. Peizerat, C. Enz (EPFL)

ABSTRACT: For a CIS readout chain based on 4T pixel, column amplification and CDS/CMS, this work confirms that thermal noise can be reduced to be neglected compared to 1/f noise. Based on analytic noise calculation and simulation results using a 180nm process, we show that 1/f and RTS noise originating from the in-pixel source follower transistor are the dominant noise sources in a classic CIS readout chain. We demonstrate that, for a given CMOS process, the input referred 1/f noise is minimal for a unique pair of gate dimensions of the in-pixel source follower and we give its expression as a function of technological parameters.

Classic low light CIS readout chains are based on 4T pixels with pinned photodiode, column level amplification with bandwidth control and correlated double sampling (CDS) or multiple sampling (CMS). Figure 1 presents the readout chain used for noise calculation and transient noise simulations.

To give an example, we plot, in figure 2, the calculated input referred noise, of a 180nm process where the minimum gate width is 0.22µm and the minimum gate length is 0.3µm, Cox = 5.10fF/µm2, cGS = 3.4fF/µm2 and cGD = 0.4fF/µm and CSN = 0.8fF.

Finally, our conclusions are:
- Thermal noise can be reduced independently from the in-pixel source follower design by means of column level gain, bandwidth control or multiple sampling.
- CMS shows no advantage over CDS for thermal noise reduction and offers slightly more 1/f noise reduction (about 20% less 1/f noise if high number of samples is used).
- 1/f noise dominates the read noise. For a given CMOS process, and a given CSN, the optimal gate dimensions W and L for the minimum 1/f noise can be obtained analytically. This transistor can be biased depending on the obtained W/L ratio.

Related Publications:
Terahertz (THz) imaging emerges as a promising field of research with many applications: medical and security imaging or non-destructive industrial testing. Furthermore major advantages are the non-ionizing characteristics of THz radiations and penetration of non-conducting and thin materials. CMOS-only THz imagers using standard process and working at room temperature tend to be a low-cost alternative to above-IC imagers (e.g. bolometer-based imagers). Based on plasma wave theory, these detectors use a MOS transistor, connected to an antenna, to convert incident THz waves to a DC voltage $\Delta U$ which is proportional to the power of the detected THz radiation.

Nevertheless readout noise is a significant issue for THz detectors which are enlightened by THz beam with low radiation power. This work tackles this problem by drastically improving the detector sensitivity. Firstly, to deal with 1/f noise, the circuit takes the possibility to modulate the active THz source that lights the scene in order to translate the base-band signal of the image around frequency $f_{mod}$ and its harmonics, allowing filtering of high flicker noise. This acts as a chopper-stabilized low-frequency noise reduction technique. The second improvement is the significant reduction of the thermal noise bandwidth realized by the in-pixel configurable signal processing chain described in Fig.1.

In addition the optimization of the couple antenna/MOS and the high gain integrated in the pixel provide a high responsivity of 300 kV/W at 270 GHz.

In addition the optimization of the couple antenna/MOS and the high gain integrated in the pixel provide a high responsivity of 300 kV/W at 270 GHz.

These performances give promising outlooks for future applications and CEA-Leti continues working on performance improvements and additional feature implementations.

**Real-time CMOS-based THz image sensor**

**Research topics:** CMOS-based THz detector, THz camera, in-pixel demodulation.

**ABSTRACT:** This research work presents a THz 31*31 pixels, 100fps image sensor fully integrated in a 130nm CMOS process. Readout noise is significantly reduced by embedding an in-pixel configurable signal processing chain. A 16-paths passive SC-filter combined with a CT Gm-C filter is integrated to filter the signal band achieving a high Q factor of 100. Thanks to these improvements this THz imager features a measured NEP of 533pW at 270 GHz and a responsivity of 300 kV/W also at 270 GHz.

![Figure 1: Bloc diagram of the in-pixel circuit for demodulation and filtering.](image)

![Figure 2: Experimental set-up and resulting images: top left: 2.5 THz image of a metallic ring held with a tape. Bottom left: 100fps sequence of a copper ruler passing in front of a 200GHz beam.](image)

**Table 1:** Summary of characteristics and comparison of the state of the art. After [1]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>533 pW @ 270 GHz</td>
<td>12 nW @ 600 GHz</td>
<td>19.8 nW @ 820 GHz</td>
</tr>
</tbody>
</table>

*1 Calculated based on the mean NEP of 28 pW/Hz$^{1/2}$ given in [3] over a bandwidth of 500 kHz, but only takes in consideration the thermal noise. *2 Responsivity including on chip gain. *3 Calculated based on the readout chain gain given in [2].

**Related Publications:**

Optical imaging through biological media is strongly limited because of light scattering. This is especially problematic in medical imaging, when the goal is to detect a millimeter-sized object within a several centimeters thick scattering medium, e.g. for early breast cancer detection. The resolution of a breast tissue image obtained from diffuse optical tomography is usually around 10 mm. As a result, emerging tumors cannot be detected. The use of an acousto-optical holographic scheme allows obtaining a 10 fold improvement resolution [1]. However, its clinical application is still out of reach because of the complexity of the setup and the limitations of the detection scheme.

In order to address these limitations, a CMOS image sensor that is dedicated to wavefront acquisition has been designed. The pixel integrates a photodiode with some analog primitives allowing the acquisition of interference patterns from which the amplitude and the phase of the light diffused in the tissues can be calculated. This CMOS image sensor also embeds a spatial phase modulation for phase conjugation and light focusing through the sample on a region of interest, for therapy. The spatial light modulator (SLM) made of liquid crystal covering the entire pixel is exposed on Figure 1. The control signals of the SLM are directly derived from the acquisition phase.

Figure 2 illustrates the 2 steps for acquisition of the interference pattern and control of the reflected light.

The architecture of the 16µm pixel has been designed in a 130 nm CMOS process. It can achieve global shutter over 4 frames, with an integration time of 250μs, or less. That architecture can then be used for 2 or 4 phases demodulation of holographic images. In order to reflect about 50% of the incident light, the pixel has a small fill factor of 10%. From a typical incident light power (red laser) of 0.5nW collected per pixel and a fill factor of 10%, 50 000 electrons are generated and integrated after 250 μs.

Such a pixel could allow great possibilities in imaging for diagnosis in living tissues, especially in skin or breast diseases.

ABSTRACT: A CMOS image sensor architecture coupling a spatial light modulator to a photodiode, for medical imaging based on acousto-optical coherence tomography, with a digital holographic detection scheme is presented. The architecture is able to measure an interference pattern between a scattered beam transmitted through a scattering media and a reference beam, on an array with 16μm pixel pitch, at 4000Hz, which is compliant with correlation time of breast tissues. In-pixel processing allows generating from the incident light, a signal to polarize an embedded light modulator used to control the phase of the reflected beam. This reflected beam can then be focused on a region of interest of a scattering media, for therapy. The stacking of a photosensitive element with a spatial light modulator on the same device brings a significant robustness over state of the art techniques such as perfect optical matching and reduced time delay in controlling light.
On Compensating Unknown Pixel Behaviors for Image Sensors with Embedded Processing

Research topics: Compressive Sensing, Image Sensor, Inverse methods

W. Guicquero, M. Benetti, A. Peizerat, A. Dupret, P. Vandergheynst (EPFL)

ABSTRACT: Some smart imagers embed image processing operations or Compressive Sensing at the focal plane level. This process introduces artifacts due to technology dispersion and unpredictable behaviors. A generic algorithm structure is presented to compensate block artifacts by appropriate post processing operations. The restoration method is composed by a three steps loop: regularize the image, fit the model parameters and update the regularization coefficient (by reweighting a fidelity term). Two specific case studies involving structured PRNU and multi-capture Compressive Sensing (CS) are presented with simulation results.

A novel approach to accurately restore an image affected by acquisition artifacts has been developed. For instance, an image obtained as a result of a general block-based processing with unknown gain and offset map can be restored. The reconstruction algorithm takes inspiration from traditional regularization methods that have recently shown remarkable results for applications such as inpainting and Compressing Sensing (CS). The algorithm exhibits a generic structure. It algorithm has been tested in two specific cases of application. First, it can help to restore an image corrupted by a high Photo Response Non Uniformity composed of a set of unknown gain masks. Secondly it provides an algorithm framework to compensate unknown non-linear behavior of the pixel integration response for image synthesis based on a multi-capture acquisition. To demonstrate its efficiency, simulations of an HDR-CS reconstruction are presented.

The generic acquisition model parameters correction algorithm (Algorithm 1) is based on a three steps loop that stop iterating if a stopping criterion is reached. Either a maximum number of iterations or a similarity metric between consecutive reconstructed images can be defined as the stopping criteria. The algorithm first initializes the signal to restore with raw input data or a zeros signal, initialize the acquisition model using the assumed parameters and assign a proper initial value for the regularization parameter.

**initialize:**
- the signal to restore;
- the acquisition model parameters;
- the regularization coefficient;

**while** Stopping criterion is not reached **do**
- Reconstruct or denoise image by regularization;
- Find the best fitting model parameters;
- Update the regularization coefficient;

**end**

**Algorithm 1:** Image restoration with partially known acquisition model parameters

Each step of the outer loop of the Algorithm 1 corresponds to a required operation. First, the signal is regularized to limit the impact of non-perfect knowledge of the model parameters relaxing constraints on the fidelity of the reconstruction according to the raw data. Then, the acquisition model parameters are optimized to fit with the previously regularized image. Finally, the regularization coefficient is updated to increase the weight of the fidelity term -- and thus reducing the impact of the regularization -- for the next iteration.

Fig. 1 shows the correction of the camera corrupted by high PRNU. This PRNU is defined by non-overlapped gain patches whose positions are set and known. This kind of artifacts models the possible effect of block based processing with different indexes for each block, producing a different PRNU for each index.

Fig. 2 presents a result on the compensation of not accurately known pixel integration responses depending on the integration time. In the case of multi-capture CS imaging with an application to HDR, a precise knowledge of the pixel behavior is required to properly synthesize the image from raw data. Fig. 2 undoubtedly shows the improvements reached when the proposed algorithm is used compared to a more naive one.

**Figure 1:** Restored camera test images corresponding from left to right to the algorithm initialization and the outputs of the 1st, 4th, 7th and 10th iteration of the outer loop.

**Figure 2:** Left reconstructed HDR image with a compensation on unknown pixel integration function. Right reconstructed HDR image without any modification of the assumed model parameters.

Related Publications:
Embedded Wire Diagnosis Integrated Sensor
For Intermittent Fault Location

Luca Incarbone, Fabrice Auzanneau, Wafa Ben Hassen, Yannick Bonhomme

ABSTRACT: This work presents an EMC (Electromagnetic Compatibility) compliant sensor able to detect and locate intermittent faults longer than 300 μs in wires in real time. The proposed System On Chip can diagnose any wire using different reflectometry-based methods. The injected signal’s spectrum is designed to ensure a continuous harmless diagnosis. The ASIC integrates the whole processing chain, i.e. the analog cells, in particular ADC/DAC, and the digital processing elements based on a custom processing unit (DSP-like) that analyzes in real time the reflected signal.

The architecture of the ASIC is presented in Fig 1. Two external communications are available: serial (SPI) and parallel (BUS). The circuit is in slave mode and an interruption is controlled by a specific unit to easily call some actions by the external master (μC).

The ASIC has three main functions: signal generation, signal acquisition and signal processing. The role of the generation unit is to drive the DAC (Digital to Analog Converter) to inject the MCTDR (Multi Carrier Time Domain Reflectometry) signal stored in the memory (Generation Memory) in a continuous loop. The acquisition unit adds the values coming from the ADC (Analog to Digital Converter) to the values stored before and records the result into the Acquisition Memory. This operation is made 32 times to get a signal with higher SNR.

Another goal of the acquisition unit is to oversample the measured signals to boost the fault location accuracy. Considering that the cable is in a stationary state (compared to the short period of signal emission ~640 ns) we can make K measurement of the same signal. For each measurement the phase of the ADC’s clock rotates by 360°/K step. All the values stored in the acquisition memory are used to recompose a composite signal having equivalent frequency K times higher than the ADC’s clock frequency.

The SoC (System on Chip) has K as a parameter written in a register; possible values are 1, 2, 5, 10, 20 and 25.

An ASIC was designed and realized for real time embedded wire diagnosis applications. On a surface of 3 mm², it integrates all analog and digital functions needed: signal generation, acquisition and processing. The fault detection and location performances are shown to be better than other TDR sensors known in the literature.

The capability to detect shorter intermittent faults is a positive feature of this architecture. The required power supply seems higher but more precise consumption measurements are under progress; it should be interesting to make a deeper analysis to identify the different sources of this current consumption.

In conclusion this work proves that the methods developed in recent years can be embedded in one single chip, with equivalent performances.

Related Publications:
Embedded Aeronautic Harness Monitoring and Reliability

Research topic: Reliability of aircraft wiring harnesses

F. Auzanneau, L. Incarbone, W. Ben Hassen

ABSTRACT: Reliability is an essential driver for avionics electronics design, and many requirements have been recommended by regulatory agencies and standards organizations. The extension to the Electrical Wiring Interconnection System (EWIS) has been recently done, leading to many researches on the use of reflectometry-based methods for the detection and location of defects in wired networks. Recent developments address the diagnosis of soft and intermittent defects, with new improved methods which can be coupled to probabilistic models, aiming finally at predictive maintenance.

Wired networks are considered as the backbones of complex systems. The increase of the complexity of modern systems has come with the increase of wire lengths (Fig. 1).

Whatever their application domain, wires can be subject to aggressive environmental conditions which may create defects. These defects can have dramatic consequences if the wires are parts of safety critical systems, i.e., systems whose failure may cause death or injury to people, loss or severe damage to equipment or environmental harm. It is therefore important to quickly detect and precisely locate failures in wired networks as it is usually done for electronic systems [1].

Recent developments are aiming at miniaturizing reflectometry sensors to embed them as close as possible to the harness. This enables diagnosis of live wires (i.e. test cables concurrently to their normal usage), using several distributed sensors to cancel location ambiguity in branched networks. Embedded diagnosis makes it possible to catch intermittent defects, such as arcs, which can last less than 1 ms but have deadly consequences.

New powerful data processing methods have shown the possibility of detecting "soft defects", i.e. defects that have only a limited impact on the wire but may be the premises of future hard defects (open or short circuits).

Figure 2: Bayesian Networks help prepare predictive maintenance and estimate the trust level of the diagnosis

Joining these techniques with probabilistic models, such as Bayesian Networks (Fig. 2), results in the increase of the detection capacity while reducing the number of sensors, thus decreasing the cost. This opens the way to predictive maintenance: optimizing the scheduling of corrective maintenance before failures happen by estimating the remaining useful lifetime of a cable or a harness.

Related Publications:
Conducted EMC measurements setup was based on the requirements of the CISPR25 norm, from the International Electrotechnical Commission (IEC). An Electronic Control Unit (ECU) from Delphi was equipped with an additional FPGA-based card implementing the MCTDR diagnosis, as shown on Fig. 1 (left). The harness was connected to the ECU, and the diagnosis signals injected into 4 different wires (signal and supply), through the ECU’s connectors. The current clamp was positioned at various places along the wires, to measure the current in the 150 kHz – 108 MHz frequency band. Various configurations were tested: diagnosis system OFF, diagnosis system ON but no signal injected, full band MCTDR (Multi Carrier Time Domain Reflectometry) injected signal and MCTDR signal with CAN and FM bands cancelled.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>CAN signals</th>
<th>Diagnosis system</th>
<th>MCTDR signals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ON</td>
<td>ON</td>
<td>No signal</td>
</tr>
<tr>
<td>2</td>
<td>ON</td>
<td>ON</td>
<td>MCTDR Full band</td>
</tr>
<tr>
<td>3</td>
<td>ON</td>
<td>ON</td>
<td>MCTDR no CAN/FM</td>
</tr>
</tbody>
</table>

Previous radiated EMC measurements had shown that the STDR method is not suited to fulfill EMC constraints. Fig. 3 shows measurement results for radiated EMC, using configurations 2 and 3. The CAN band is too low to appear in this measurement setup, but the cancellation of the FM band is clearly visible. A reduction of 20 to 30 dB on the emitted fields has been achieved.

Conducted and radiated EMC measurements have shown that cancelling frequency bands in the diagnosis signal allows fulfilling automotive standards. This work confirms that, from the EMC point of view, MCTDR is the only wire diagnosis method that can be used for embedded wire diagnosis.

Related Publications:
Self-Adaptive Correlation Method for Soft Defect Detection in Cable by Reflectometry

Research topics: Reflectometry, Soft defect, Cable diagnosis

S. Sallem and N. Ravot (CEA LIST)

ABSTRACT: The self-adaptive correlation method (SACM), based on a temporal processing, consists in amplifying the signatures of soft defects and make them more easily detectable. The algorithm amplifies any signature correlated with the injected signal. This method allows to highlight soft defects while attenuating the noise present in the reflectogram. The term "self-adaptive" means that the gain is automatically adjusted depending on the amplitude of the soft defect signature. This method is applicable to different reflectometry techniques with different waveforms and different position of the defect in the cable.

The advantage of the self-adaptive correlation method (SACM) versus time-frequency methods [1],[2] (Wigner-Ville) consists in its computational simplicity (low cost). In addition, it avoids false alarms (the cross-terms) that limit the methods based on Wigner-Ville transform. In SACM, cross terms don't exist. We tested this method on different types of cables (twisted pair AWG24 and coaxial cable RG58CU) by simulation and experimentally.

Application of the SACM to TDR:
We inject a temporal Gaussian pulse into the cable extremity. A small temporal width pulse is required to enhance the detection of a small soft defect, but we are limited by the frequency bandwidth.
We tested the proposed method on a five meters shielded twisted pair cable AWG24 by simulation and experimentally.
The soft defect consists in removing the sheath and the shielding of a small portion of the cable (about 2cm in the middle of the cable), Fig. 1. This type of cable (shielded twisted pair) is widely used in the aeronautic industry.

Fig. 1 AWG24 twisted pair with a soft defect

We tested the proposed method on a five meters shielded twisted pair cable AWG24 by simulation and experimentally.
The soft defect consists in removing the sheath and the shielding of a small portion of the cable (about 2cm in the middle of the cable), Fig. 1. This type of cable (shielded twisted pair) is widely used in the aeronautic industry.

Application of the SACM to transitions:

We consider a point-to-point line consisting of three sections of coaxial cable RG58CU, connected by two transitions. The reflectogram obtained by the injection of a gaussian pulse (1ns width) to the input of the line was processed with the SACM algorithm. We find that we can detect and locate transitions on the line very clearly as shown on fig. 3. This allows to monitor changes in the junctions state and anticipate any damage.

Fig. 3. Application on a coaxial point-to-point line with two transitions

Weaknesses in wired networks are often the result of transitions arising from connecting/disconnecting cable sections. It is therefore important to monitor these sections in a permanent way.

We consider a point-to-point line consisting of three sections of coaxial cable RG58CU, connected by two transitions. The reflectogram obtained by the injection of a gaussian pulse (1ns width) to the input of the line was processed with the SACM algorithm. We find that we can detect and locate transitions on the line very clearly as shown on fig. 3. This allows to monitor changes in the junctions state and anticipate any damage.

Related Publications:
Although reflectometry has proven its efficiency in fault location, it may suffer from ambiguity in branched wiring networks. Distributing diagnosis systems (reflectometers) is a possible way to overcome this problem. However, the presence of influencing factors on reflectometry measurement may also impose uncertainties related to the diagnosis reliability and then the decision-making.

As solution, we propose a probabilistic approach based on Bayesian Network (BN) for wiring network diagnosis. The main objective is to evaluate the diagnosis reliability by considering influencing factors on reflectometry measurement in complex wiring networks. The proposed approach is described in Fig. 1. It includes two major steps which are: local diagnosis and global diagnosis.

1- Local diagnosis: Cable diagnosis reliability: In order to reduce the system complexity, a BN is modelled for each cable in the network where knowledge related to the cable life profile (or also called influencing factors) are introduced and propagated in the BN to calculate the conditional probability of the presence of the fault on the considered cable. In this paper, the selected influencing factors are related to the environment (thermal and mechanical stress), cable characteristics (aging, type, length, bandwidth), reflectometer reliability in signal emission and reception, etc. Figure 2 presents the model of BN. The causal relationships between the influencing factors are established based on experimentations, simulations, expert opinions, etc.

2- Global diagnosis: Data fusion
The probabilistic information obtained by each reflectometer on each cable is aggregated to calculate the global probability of the fault location on each branch of the studied network. The data fusion may be applied by using a weighted sum where a weight is accorded to each reflectometer based on its state, capacity, performance, location in the network, etc. Then, the decision about the fault location is made based on the data fusion results. In fact, the data fusion permits to increase the diagnosis reliability even in the case of deficient reflectometer(s). To do so, the Orthogonal Multi-Tone Time Domain Reflectometry ensures communication among reflectometers [2].

Rather than diagnosis reliability evaluation (sense factors-results), the modelled BN may be used in the sense results-factors to determine the probable causes of the appearance of the fault on the network. Moreover, the introduction of the time in the BN (dynamic BN) permits to estimate the cable life time (prognostic).

Related Publications:
Wireless Communications & Cyber-Physical Systems

Ultra-Low Power Radio
Millimeter Wave Radio
Antenna tuning
Energy Management of Sensor Networks
Middleware & Cyber-Physical systems
Continuous Time Digital Signal Processing for Ultra Low Power Radios

Research topics: CT-ADC/DSP, Ultra-Low-Power, Wake-Up Radios

A. Ratiu, D. Morche

ABSTRACT: Continuous time digital signal processing represent a class of systems in which the information is encoded in a set of discrete levels in the amplitude domain as well as in the timing between the arrival of these levels. These systems present a set of interesting properties (activity-dependent power dissipation, require no clocks, good programmability) which we try to exploit for the design of an ultra low power radio back-end. The application targeted by this design is the implementation of an always on wake-up radio for wireless sensor networks.

CT analog-to-digital conversion can be viewed as defining a set of discrete digital levels and triggering a digital output whenever a level is crossed by the input signal (Figure 1). The information is thus encoded in the digital level as well as in the time of the level crossings. The output signal can be processed by a continuous time DSP made up of async delay cells and async adders which preserve the timing between events (Figure 1). Such systems have been proven to be energy efficient in performing interferer rejection for ultra-wide band signals [1]. In this design, however, we are focusing on an ultra low-power implementation suitable for always-on radios.

Figure 1: Working principle of a CT-ADC (left); view of as simple CT finite impulse response filter (right)

One challenge in designing CT-ADCs/DSPs is the fast simulation of such systems. The CT-DSP can be easily simulated using an event driven approach in which level crossing events from the CT-ADC propagate from one system block to another with a certain behavioral and time delay model. However, the efficient generation of the level crossing events at the CT-ADC level is problematic since it is difficult to obtain a high precision on the time of the level crossing while using a fast simulator.

Analog simulation is precise for device level simulations but too slow for architecture exploration. On the other hand, discrete time simulation provides a fast and simple way of simulating CT-DSPs but requires a simulation oversampling ratio (OSR) much greater than the highest frequency present in the system. A low OSR can introduce a simulation noise greater than the noise in the system thereby rendering the simulation useless.

In [2] we have proposed a hybrid simulation method where we use a discrete time simulation coupled with a first order interpolation in order to efficiently generate the level crossing events. A theoretical model has also been developed for the proposed simulation method enabling us to predict the simulation error based on some simple characteristics of the input signal. This theory has been validated by simulations with simple input signals (single sinusoids) as well as signals from two wireless standards (802.11a and 802.11b).

As seen in Figure 2, for a simulation noise level of 30dB the proposed simulation method requires an OSR of 32, compared to 1024 required by a discrete time simulation. This reduction in OSR represents a significant gain in the simulation time, enabling a thorough architecture search.

The proposed simulation method has been used to study the impact of several transistor level non-idealities of different blocks of the CT-ADC [3] presented in Figure 3.

Results presented in [4] showcase the amount of non-ideal behavior acceptable before different CT-ADC performance metrics start to degrade. This enables us to optimize the power consumption of each block without sacrificing the overall performance of the conversion.

Currently, we are working on the implementation of a low power, intermediate frequency stage for rejecting interferers using the previously described principles.

Previously we have designed, layouted and post-layout simulated a 15-phases single-ended RO in FDSOI 28nm, optimized for a high bandwidth PLL for Bluetooth Low Energy (BLE) applications. Specifications are as follow: Power consumption: 0.8mW, Phase noise: -90dBc@1MHz at FOSC of 2.45GHz. From this VCO, with MonteCarlo simulation, we extracted a standard deviation (std) of the phase propagation delay, of about 5ps [4]. A VCO phase noise (L(f absolute)) of -90dBc@1MHz (Df=1MHz) gives a jitter of J=260fs RMS. So, a maximum mismatch of 260fs RMS between phases has to be reached in order to assume a spurious power lower than -90dBc@1MHz.

Spurious simulation in a complete PLL circuit at transistor level can be done using transient simulations and K. Kundert Matlab code. PLL phase noise is computed from the Fourier transform (FFT) of 1-million period-values vector, extracted from transient simulation. We used this method to get a reference PLL phase noise chart (without spurious) to validate our phase noise model (Fig. 4). We can notice that the total PLL output phase noise is -95dBc/Hz@1MHz and matches the VCO phase noise.

A Matlab model of our FPD PLL schematic (fig. 1) has been created to calculate spurious content from the deterministic jitter created by FPD divider. Using this method, spurious simulation time for FPD architecture is greatly decreased, as there is no need for long transient simulations to get spurious characteristics (number, frequencies and power) for each channel, i.e. divider phase shifts. In the same way, if we measure the spurious of the PLL, it is possible to determine, with this code, the value of the PLL phase noise.

PLL blocks (REF, PFD (phase frequency detector), CP, LPF and VCO) phase noise contribution are evaluated and added to compute the total phase noise without spurious. After, we can compare the spurious, with the total PLL phase noise.

Spurious simulation with 5ps RMS phase mismatch shows that the adjacent channels spurious power is -65 and -75 dBc/Hz at 2MHz and 4MHz (respectively).

We can now specify the ring oscillator which includes a low mismatch depending of the phase noise specification of the desired application.

To design a very low power PLL for Bluetooth Low Energy, we have chosen FPD architecture for its simplicity and power efficiency. To verify feasibility of large bandwidth PLL with low power ring oscillator, we have developed phase noise and spurious models. With these models of spur generation, we accelerate simulation time and can predict spurious power and frequency from extracted phase mismatch, and determine VCO phase mismatch specifications.

As we show in our simulation, the spurious generated by phase mismatch of the optimized 15-cell VCO are above the PLL phase noise. The next step of this work is the design of very low mismatch inverters for ring oscillators whose spurious comply with the phase noise specifications.

Related Publications:
SOI CMOS Tunable Capacitors for Antenna Tuning

**Research topics:** SOI CMOS, Tunable Antenna, Integrated RF Front-End

A. Giry, D. Nicolas, E. Ben Abdallah, S. Bories

**ABSTRACT:** The need for wireless devices with small form factors is driving research towards reconfigurable miniature antennas. In this research work, tunable capacitors have been designed and integrated in a SOI CMOS technology to allow aperture tuning of a small notch antenna. The proposed frequency-agile antenna can be tuned from 420MHz to 790MHz, with an associated radiation efficiency of 33% to 73%, compared to a 5% to 62% radiation efficiency without antenna tuning.

Since the first commercial developments of cellular communications, multiple parts of the spectrum have been absorbed by the ever growing demand of cellular systems. In the sub-GHz range more than 10 different frequency bands have already been defined by LTE-A standard and further bands will be added in the future as more spectrum will become available worldwide for next generation mobile communications technologies. Wireless devices with small size lead to serious challenge for multiband antenna design operating in this sub-GHz range since miniature antennas naturally present a narrow bandwidth. To address these conflicting requirements of size and frequency range, one solution is to add a tunable component to the antenna structure to allow adjustment of antenna’s electrical length (aperture tuning).

In this work, tunable capacitors have been designed and implemented in a SOI CMOS technology to provide aperture tuning to a small notch antenna [1]. By placing the tunable capacitors on the radiating element, frequency tuning of the antenna is obtained which allows to address multiple frequency bands ranging from 420MHz to 790MHz. SOI CMOS technology, which provides low ON-state resistance and OFF-state capacitance switch devices, in addition with high-Q inductors and capacitors, represents a cost-effective solution for low-loss high-power tunable capacitors integration [2].

As shown in Fig.2, the proposed tunable antenna consists of a slot antenna realized on a FR4 substrate and a SOI CMOS die, including two tunable capacitors and a SPI control interface, inserted at the end of the slot.

---

**Related Publications:**


ABSTRACT: A new concept based on the injection of a periodically repeated oscillations train into an oscillator is presented. It is based on a phenomenon named pseudo-locking, leading to a novel programmable harmonic selection technique. Theoretical description of the concept is presented, in particular an analytical condition is carried out, paving the way of design rules. For a proof of concept, an oscillator fabricated on a 40 nm CMOS technology was measured. The concept was validated by measured phase noise equivalent to classical fundamental injection locking phase noise performance.

In this work, an original use of an injected oscillator is presented. It consists in injecting into the oscillator a Periodically Repeated Oscillations Train (PROT). In the time domain the PROT signal consist of a 1/FPRP-periodic alternation of oscillations (on states) and off states. In the frequency domain the PROT signal corresponds to an FPRP spaced multiharmonic signal with a spectral envelope that is a cardinal sine centered at \(\omega_{\text{inj}}\). The PROT signal is generated by switching ON and OFF an oscillator with a TPRP-periodic signal. This signal can be used to perform a large factor frequency multiplication, but it can be only applied if the harmonic of the wanted frequency can be selected. This is accomplished by the injected oscillator. Figure 1 shows the operation principle and the PROT signal characteristics. The injected oscillator uses the PROT signal which is TPRP periodic. When the injected oscillator achieves TPRP periodicity on its output phase transient behavior, a new phenomenon is observed called “pseudo locking”. Under this condition, the resulting sine-wave signal at the injected oscillator has spectral performances (frequency and phase noise) that depend only on the injection PROT.

The verification of the theoretical operation has been done with a 60 GHz PROT generator and injected oscillator fabricated in a 40nm CMOS process. Measurements were performed on both stand-alone circuits and also when the PROT signal is injected to the injected oscillator. Figure 2 summarizes the main results on both signals in time domain and frequency domain. The PROT signal is clearly a 60 GHz band oscillation that switches on-and-off with TPRP period. The PROT spectrum consists on a FPRP spaced multiharmonic signal. The output signal is a sine-wave signal whose frequency is exactly an integer multiple of the FPRP frequency (29*FPRP). The output signal spectrum includes only the selected harmonic (62.64 GHz) of the PROT signal. Phase noise measurements show that the ILO reproduces the PROT signal phase noise all over the frequency offset. Hence, the ILO phase noise behavior corresponds to a fundamental injection locking and this system can be seen as a high order micro-wave regenerative filter. Moreover, the PROT generator combined to the injected oscillator leads to new techniques for high order efficient programmable frequency multiplication.

![Figure 1: PROT signal generator, injected oscillator and resulting output signal.](image1)

![Figure 2: BAW filter Equivalent Load response for classical and new frequency offset.](image2)

Related Publications:
Testing every performances of an analog RF circuit is an expensive task, especially in ultra narrow band signals when bit error rate extractions are needed [1]. Indeed, it requires using expensive sensors and can take long time to obtain accurate measurement. Moreover, with the dimensions scaling, technology variability is continuously growing. Its influence on circuit performances can be reduced by increasing the size of the elementary devices (when possible) but this translates automatically into higher silicon area and more power consumption which is unacceptable. As a consequence, having wide margin with respect to performance dispersion is not compatible with low cost, low power SoC. Thus, it has to be judiciously optimized [2], leading to some out of specifications components. Detecting out of specifications devices makes the test more difficult than simply detecting non-functional ones. Circuit tests are usually carried out via the direct measurement of performances. Due to the multi-standard requirements, SoCs are getting more and more complex thus a wide variety of functionalities has to be validated. Consequently, the complexity of the Automatic Test Equipment (ATE) is also increasing, leading to a higher cost. In the end, test cost has an important impact on final product price. Therefore, the motivation to reduce the test duration is high.

The alternative measurements strategy aims at replacing the most expensive performance measurements by the measurement of internal circuit signals in order to decrease the test cost while maintaining the test coverage. Such a strategy requires an extra analysis step, see Fig.1. A well-known method for Alternative Measurement is the Built-In Self Test (BIST) method. The main challenge in alternative measurements based test is to correctly choose the intern signals to be measured so that the test coverage is correct, and so that the overall test cost is low enough.

This balance is usually equilibrated by test engineer experience. This work proposes a novel methodology, based on data-mining, to address this problem, as depicted in Fig.2.

The method is built using data mining algorithms, and especially the CDC-Focus algorithm [2], which allows to extract from simulations or measurements the more relevant and cheapest subset of test steps. Results obtained allow selecting the required tests which will lead to specified test metric (i.e. false-positive ratio, test coverage…) using several classification models: MARS (Multivariate Adaptive Regression Splines), SVM (Support Vector Machine), kNN (k-Nearest Neighbors). Those results allow selecting tests to balance test cost and accuracy of test sequence.

**Related Publications:**


Differential noise figure measurement using the correlation of output noise waves

Research topics: noise measurement, differential amplifiers, network analyzers

Y. Andee, J. Prouvée, A. Siligaris, F. Graux (Rohde&Schwarz), F. Danneville (IEMN)

ABSTRACT: Our research work deals with the noise figure measurement of differential amplifiers using 4-port network analyzers. The approach is based on the determination of the correlation of output noise waves in terms of the 4-port S-parameters and of the output noise powers. The measurement setup is simple as it does not require any hybrid coupler or calibrated noise source. Measurement results of an on-wafer differential LNA demonstrate the validity of our new approach.

Differential circuits are increasingly designed for radiofrequency and other high-frequency applications, taking advantage of their immunity against common-mode noise and interference. The noise figure measurement of differential circuits is a challenging subject, particularly due to the correlation of output noises. In literature, noise figure measurements of differential circuits are performed using 180° hybrid couplers. This approach allows the de-embedding of the differential noise figure without the need of measuring the correlation of output noises. However, it makes use of couplers which are bandwidth-limited and it requires a proper de-embedding which is quite complex. Our research work deals with the noise measurement of differential amplifiers without the need of couplers. It is based on the determination of the correlation of the noise waves at the output ports of the differential amplifier.

An analytical expression of the differential noise figure of a 4-port device is obtained in [1]. This expression is derived from the noise-wave formalism, illustrated in Fig.1., and the mixed-mode S-parameters. The expression of the noise figure is valid for a system where the sources and the loads are well matched. The differential noise figure is given by:

\[ \Gamma_{\text{diff}} = \frac{b_3^2 + b_4^2 - 2 \text{Re}(b_3^* \cdot b_4^*)}{2kT \Delta f \left( S_{\text{S}21}^2 + S_{\text{S}22}^2 \right)} \]

where \( S_{\text{S}21} \) and \( S_{\text{S}22} \) are the mixed-mode gains which can be calculated from the classical S-parameters.

\[ b_3^2 + b_4^2 = \frac{S_{\text{S}21}^2 + S_{\text{S}22}^2}{S_{\text{S}31}^2 + S_{\text{S}32}^2} \]

\[ b_3 = \frac{S_{\text{S}21}}{S_{\text{S}31}} \cdot b_1 - \frac{S_{\text{S}22}}{S_{\text{S}32}} \cdot b_2 \]

\[ b_4 = \frac{S_{\text{S}21}}{S_{\text{S}31}} \cdot b_1 + \frac{S_{\text{S}22}}{S_{\text{S}32}} \cdot b_2 \]

Based on this theoretical work, a fast and functional method for the noise figure measurement of an on-wafer RF differential amplifier is developed using a 4-port Network Analyzer [2]. The technique consists of connecting the differential DUT to the analyzer test ports via 2 dual probes. The input ports are connected to two independent internal 50Ω sources of the analyzer and the output ports are terminated with two internal 50Ω receivers. An SOLT calibration procedure is performed and the 4-port S-parameters and the mixed-mode gains of the amplifier are directly obtained. The noise powers at the output ports are measured using RMS detectors after receiver power calibrations have been performed. Using the S-parameters and the noise powers, the correlation is calculated and the differential noise figure is obtained as shown in Fig.2.

Figure 2: Differential noise figures measured with our technique and with the coupler method

The agreement between the 2 noise figures confirms the validity of our new approach. The advantage of this technique is that it requires a simple measurement setup where no coupler or noise source is required.

Related Publications:
Wireless transceivers must perform adequately when the channel conditions are in worst case. However, when signal and interference conditions are favorable, the transceiver may consume more energy than required. Therefore, a reconfigurable transceiver, able to adjust its corresponding power consumption to the time-varying signal and interference conditions, could have a decisive impact on the lifetime of the sensor node.

For example, in Fig. 1(a), the desired signal is weak and the adjacent channel interferers are strong. The receiver must therefore operate in high performance mode which minimizes noise floor and phase noise while selectivity, linearity, and dynamic range are maximized. In (b), the moderate signal and interferer levels imply that the receiver's performance levels can be relaxed, hence resulting in power consumption savings. In (c), the strong signal and weak adjacent channel interferers imply that the receiver specifications can be further relaxed thus providing a maximum of power savings.

In this work, a simulation framework for wireless sensor networks (WSN), EnvAdapt, is presented in which the modeling of reconfigurable transceivers is made possible. Using this tool, different reconfiguration strategies can be examined and real-life WSN scenarios can be considered to show that transceiver reconfiguration can help reduce power consumption, thus extending the lifetime of the sensor node.

In addition, the framework has been specially designed to test novel intra-frame reconfiguration algorithms as well as cross-layer link quality estimation techniques. Thus, complex WSN scenarios can be built that test the applicability of reconfigurable hardware components in order to spur their development. The framework can also be used to evaluate how to design next generation upper layer communication protocols that best exploit the transceiver's reconfiguration capabilities.

Thanks to this novel tool, real-life WSN scenarios including detailed channel, mobility, and application models can be built to show that transceiver reconfiguration can help reduce power consumption, thus extending the lifetime of the sensor node.

Related Publications:
Dynamic Power Management at Sensor Network Level using Model Predictive Control

Research topics: Wireless Sensor Network, Model Predictive Control, power control

O. Mokrenko, S. Lesecq, W. Lombardi, D. Puschini, C. Albea (LAAS), O. Debicki

ABSTRACT: Technological advances have made wireless sensor nodes cheap and reliable enough to be brought into various application domains. These nodes are usually powered by battery, thus they have a limited lifespan. This work addresses the control of power consumption of sensor nodes equipped with batteries at the network level. Power management is used to dynamically re-configure the set of nodes in order to provide given services and performance levels with a minimum number of active nodes. The power control formulation is based on Model Predictive Control. Simulations are performed to demonstrate the efficiency of the proposed control method.

Wireless Sensor Networks (WSNs) are usually made of low-power, low-cost, and energy-constrained sensor nodes (SNs) that are deployed to sense a physical phenomenon of interest. The measurements are collected by access points. From these measurements, an application is built to monitor and/or control the physical world. In both situations, the nodes can be dedicated to specific roles, namely, sensing, actuation and control, or to any combination of these main roles. Whatever the assigned functionality, the node embeds processing capability associated with memory capacity, both possibly limited.

Consider a WSN that contains \( n \) sensor nodes \( S_i \), \( i = 1, ..., n \), powered by batteries. Each SN can be placed in different functioning modes \( M_j \), \( j = 1, ..., m \), which are related to the working state of the node, characterized by a known average power consumption over a given period of time. Here, the nodes are supposed to provide redundant information. Thus, if one selects properly the mode of the nodes, power savings can be achieved.

In order to control the energy consumption in a WSN, the remaining energy in the batteries of the nodes is modeled with a discrete-time state-space model:

\[
x_{k+1} = Ax_k + Bu_k
\]

where \( x_k \) is the remaining energy in the battery of the nodes \( S_i \) at the time instant \( k \). \( Bu_k \) represents the averaged energy that will be used during the time interval \([k, k+1]\). Moreover, \( u_k \) takes its values in \([0, 1]\)\(^m\) (i.e. binary control value).

In order to fulfill a “mission” while decreasing the power consumption, a subset of \( d_j \), SNs is assigned to a given functioning mode \( M_j \). Thus an extra set of constraints is defined:

\[
\sum_{i=1}^{n} u_i(j) = d_j
\]

A dynamic energy saving control policy has to be implemented at the WSN level in order to increase its lifespan while guaranteeing adequate performance levels expressed as a mission. However, it imposes extra functional constraints to the WSN.

The minimization of the power consumption of (1) can be presented as a Quadratic Programming (QP) problem:

\[
u^* = \arg\min_u \sum_{i=0}^{N_p-1} x_{k+i|k}^T Q x_{k+i|k} + \sum_{i=0}^{N_u-1} u_{k+i|k}^T R u_{k+i|k}
\]

subject to

\[
\begin{align*}
x_{k+i+1|k} &= Ax_{k+i|k} + Bu_{k+i|k}, & i = 1, ..., N_p - 1 \\
u_{k+i|k} &= 0, & i = N_p, N_p + 1, ..., N_u - 1 \\
X_{\min} &\leq x_{k+i|k} \leq X_{\max}, & i = 1, ..., N_p - 1
\end{align*}
\]

where \( Q = Q^T \geq 0 \) and \( R = R^T \geq 0 \) are weighting matrices, \( X_{\min} \) and \( X_{\max} \) are the lower and upper bounds on the state \( x_{k+i|k} \). The prediction horizon \( N_p \) the control horizon \( N_u \leq N_p \) together with matrices \( Q \) and \( R \) are the degrees of freedom of the control design methodology.

In the application case, when the Dynamic Power Management (DPM) at network level is not implemented, the network lifespan is equal to 302 hours, while it is equal to 700 hours with the DPM technique. This means that, for this particular scenario, the proposed management technique doubles the lifespan. Figure 1 compares the total energy consumption in both situations (i.e. with and without DPM).

**Figure 1: Total energy evolution comparison (with and without DPM)**

Related Publications:
Coordination in the context of the connected objects.

Research topics: Coordination, Rapid Prototyping, Demonstrators

M. Louvel, F. Pacull, M.I. Vergara-Gallego

ABSTRACT: Coordination of systems may rapidly become a daunting task without using the proper mechanism. LINC is a compact yet powerful coordination environment for building applications combining distributed software and hardware components. Our approach is to leverage on the different existing initiatives in IoT and by proposing a glue layer to coordinate them. Coordinating is not just making them talking together, it is also question of making them behaving in a coherent way to reach a common goal. This is a real challenge for IoT

Today's systems are not only distributed, they are composed of other systems more or less opaque. They have to interact with real world and thus have to consider on the one hand very small embedded systems and on the other end unbounded resources geographically sprayed. We have developed with LINC, a coordination environment which provides a layer on top of such networks that does not make it reliable (we know for at least 3 decades that it is impossible) but that enforces some guaranties and properties. This simplifies the task of the application or service programmers by providing a clear semantic concerning the potential failure and a mean to deal with them in a coherent way. The core concept of LINC and its relationship with earlier environments can be found in [1]. LINC offers an abstraction layer based on distributed tuple-spaces (called bags) providing a shared associative memory. Three primitives allow to read, get and put tuples into the bags. This bag abstraction is a very convenient way to see in a uniform way a database, physical world (e.g. sensors or actuators) and legacy systems. These primitives are used through coordination rules containing a precondition and a performance part. Precondition verifies the presence of resources in the bags. These resources model in a uniform way conditions, events, or status and define when to trigger the performance part. The latter embeds in transaction a set of actions which consumes, modifies and insert resources in the bag. Atomicity (all-or-nothing) at the level of the set of actions prevents the system to fall in an inconsistent state where only part of the actions has been done. In such a case it would be required to unroll code to try to reach an improbable stable state. With LINC by construction you ensure that either you cross the river or you stay on the initial bank in a stable situation to envisage an alternative solution. You will never be in the middle of the stream, trying to painfully reach one of the two banks. Through demonstrators we have shown that LINC environment can fit several targets both in term of application domains and size of the systems (largely distributed or embedded). Several demonstrators have been built in the frame of the IRT Nanoelec in the domain of the silver economy and connected transport. We also did several demonstrators in collaborative projects. In ARROWHEAD (energy efficiency of industrial systems), LINC helped to deploy a system optimizing the lift energy consumption and harvesting. It coordinates a virtual lift in Valence (Sodimas), an energy optimization system in Grenoble (Schneider) and the logic to exchange only the required level of information at CEA. The system (Figure 1) offers a web interface accessible from any computer to control a cross-LAN coordinated distributed application. In FP7 SCUBA (building automation system BAS), we controlled the automatic reconfiguration of a room equipped with LON devices (lights, HVAC, push buttons ...) according to the presence or not of a separation wall (in red in the Figure 2). Here LINC is not only used for intermediation where heterogeneous devices are coordinated together by software but also for reconfiguring the hardware controllers of a given technology (here LON). Thus, once the reconfiguration is completed, LINC is not required anymore and the BAS can run autonomously. Another point is the direct encapsulation of the legacy tool used for configuring by hand the LON system (requiring several hours of a skilled engineer) in such a way a LINC rule may do the job in a couple of seconds. This rule can be triggered by a wall detector, a remote controller or even information coming from the room occupation agenda. Finally we have shown in [4] how we can build complex tangible interfaces: 384 RFID readers embedded in a table, multiple inputs and outputs including a 2D and 3D engine and 3D mouse).

Related Publications:
Challenges and New Approaches for Dependable and Cyber-
Physical System Engineering (DeCPS)

Research topics: Cyber-Physical Systems, Dependability, Real-Time, Certification

Daniela Cancila, Jean-Louis Gensternayer (CEA) Charles Robinson, Laurent Rioux (Thales)

ABSTRACT:
The authors organized the June 2014 workshop DeCPS as satellite event of the 19th
International Conference on Reliability Software Technologies - Ada-Europe. This paper
highlights the strategic overview underlying workshop’s communications and the discussion
emerging during the workshop roundtable. The success of this workshop and the industrial
feedback suggest consideration for a follow-up workshop for 2015.

In recent years, we attended a crescendo of industrial and research interest in Cyber-Physical Systems (CPS). One
distinguishing trait of CPS is that they integrate software control and decision making with signals sensing to enable
actions in and on an uncertain and dynamic environment. CPS often involves heterogeneous systems. Their design
makes extensive use of interfaces and models (tools, systems, languages) and Horizon 2020 program framework
of the European Union devotes considerable attention to various aspects of the CPS challenges. Similar trends exist
in the EIT ICT Labs, Knowledge and Innovation Communities set up by the European Institute of Innovation
and Technology to drive European leadership in ICT innovation for economic growth and quality of life.

The final roundtable led by J.L. Gerstenmayer revealed four main challenges:
1. Handling the impact in the separation of functional and non-functional attributes to meet correctness-by-construction methodology
2. Choosing which formal methods are suitable to deal with dependability, included modular certification, in industrial applications
3. Reconciling differences in the semantic interpretations of the same diagram, and
4. Facing the heterogeneity of languages, tools, teams and knowledge.

Hereafter, factors which leaded the debate are introduced. The separation of functional and non-functional attributes
has been strongly promoted by the academic and industrial communities to improve reuse of components. In many
industrial systems, this separation is a refined practice for the engineering of industrial systems. The ever increasing
complexity of systems and materials, which require more and more performance, suggests first to deploy functionality
with different levels of heterogeneity (for example in safety levels or temporal attributes) on the same material, and,
secondly, to exploit material mechanisms to ensure safety and security properties. The control of a system’s behavior
to activate the appropriate safety measures/mechanisms in the event of errors occurring involves the verification and
the control of non-functional attributes, including the real-time-related ones. It is a particular necessity for cyber-
physical systems which involve critical aspects. The scientific and industrial community agrees with the
following sentence: control of real-time properties is the cornerstone to achieve the correctness-by-construction
approach to design. However, real-time properties are often specialized and analyzed later in the design phase. The main
reason is that they are related to the adopted material, which can change, by virtue of the use of newest technologies.

This complex and often contradictory industrial context (cost / novelty) provided the core of the workshop discussions
from the first issue. Without the ambition to settling the issue during the workshop-day, we agreed with the importance
to devote effort in structuring the relationship between software and material requirements and sw and hw
mechanisms to ensure safety properties.

Related Publications:
In the last decade, Cyber-Physical Systems (CPS) have assumed an increasingly significant role in a number of disciplines, especially in Computer Science, and form one of the cornerstones of the study of dynamical and heterogeneous systems. CPS combines signals from physical components with (embedded) software components and integrated circuits (See Figure 1).

This work arises from the FSF project (Fiabilité et Surêté de Fonctionnement Reliability and Safety). The bulk of the FSF project deals with safety-related properties of a railway system that involves components, which have an inherent different nature and, to complicate the scenario further, combine different safety integrity levels (SIL).

The considered case study is in the scope of the Communication Based Train Control (CBTC) system, and considers more precisely a subset of the Automatic Train Control subsystem (ATC). The associated operational scenario is the following: a train stops at a station that is equipped with a physical barrier and automatic doors, whose purpose is to protect passengers from the moving train (see Figure 2). In order to be able to operate train and platform doors, the doors of the train and the doors of the platform need to be aligned. At that point, both of them are automatically opened - thus allowing the passengers to get on and off the train.

We refer to this phase with the technical term "passenger exchange". Finally, the train is authorized to move on if and only if both platform and train doors are closed.

Safety issues have a prominent role, especially in those CPS which ought to entail a certification process. This is exactly the case of some functionality and mechanical components of our use case.

Contract-based approaches are considered as a promising means to deal with CPS. A contract is a pair (assumption, guarantee), where the guarantee specifies the functionality provided by a component to the environment; and the assumption sets forth the conditions required from the environment in order for the component to accomplish its guarantee [2]. The contracts, which are specifications on both physical and computational components, help us identify precisely the conditions for a correct interaction.

The outlined vision exploits contracts as a means to identify precisely the conditions for a correct interaction of components as well as to specify which assumption a functional level (code) should require to a hardware level to ensure the acceptable threshold of SIL.

The industrial feedback of our result is promising and strongly positive.

Related Publications:
Optimized hardware IPs
FD SOI circuits & strategies
Thermal Management
3D ICs
Delay Monitoring

Energy, Delay & Thermal Optimized Digital Design
A Subsampling-Based IEEE 802.15.4 Demodulation Processing 
Targeting Self-Adaptive DBB for Energy-Efficient WPANs

Research topics: WPANs, Ultra-Low-Power, Digital Baseband

V. Lenoir, D. Lattard, A. Jerraya

ABSTRACT: In the context of the Internet of Things, the design of an highly-integrated wireless transceiver is a challenging task due to the limited energy budget assigned to it. The main issue lies in the variability of the radio channel which is not managed by the demodulation processing, resulting in a poor energy-efficiency. Thereby, adaptivity is a key technique to address the need of low-power consumption in unpredictable environments, by providing at any time a low-margin operating mode. Thus, we have proposed a subsampling-based IEEE 802.15.4 digital baseband enabling a tunable performance vs. energy trade-off to minimize the power consumption.

The IEEE 802.15.4 physical layer for the 2.45 GHz ISM band combines a Direct-Sequence Spread Spectrum (DSSS) technique with a digital phase modulation (O-QPSK). In summary, each transmitted byte are divided in 2 symbols, which are then associated to their corresponding pseudo-random and nearly-orthogonal sequences of 32 chips (modulation bits). At the receiver side, the baseband processing consists mainly of non-coherent detection of these DSSS using correlations which are exploited for the signal synchronization and the frame decoding.

Basically, the correlation results in a high value in case of identical and tightly aligned sequences and a near-zero value otherwise. This allows the recovering of the symbol number and the transmitted bits by testing all the possible DSSS. Nevertheless, the correlation output level depends on many variables, including the signal and noise energy at the input but also the synchronization quality. Indeed, these factors have a direct impact on the orthogonality and consequently on the mis-detection rate. Therefore, the sequences length, i.e. the quantity of chips, mitigates these issues by offering a sufficient correlation gain to maintain the required Bit Error Rate (BER). The trick is that this margin is most of the time not justified, for instance if the transmitting distance is very short. Therefore, our solution consists to randomly subsample the sequences upstream to the correlation stage, decreasing the baseband computation load [1] (the randomness allows to de-correlate the deleted chips from the sequences structure). In practice, this tunability knob causes a reduction of the correlation gain and higher cross-correlations between different DSSS. Thus, it offers several operating points to achieve a same requirement, here the BER, as illustrated in Figure 1.

For a hardware point of view, the digital baseband is built around a variable-length DSSS correlator whose segments can be disabled individually using clock-gating. Thus, it is filled according to the subsampling rate applied upstream in the datapath [2]. The DBB has been implemented in a 65 nm Low Power CMOS technology. The total gate count after synthesis is approximately of 22.8 kgates. As a reference, a standalone version of the digital baseband has been synthesized and the gate count overhead is of 2 %. Finally, after backend the obtained core area is around 0.173 mm². The power consumption of the design has been estimated with the Synopsys Primetime tool using back-annoted netlist and activity extraction from logical simulations. In this setup, several operating points have been characterized, corresponding to different rates ranging from 75 % to 0 %. The results demonstrate a power consumption scalability from 208 µA down to 82.9 µA which is at the state-of-the-art while offering a potential use in a self-adaptive wireless transceiver. As an illustration, the Figure 2 gives the power consumption profile of the digital baseband during a packet demodulation where the payload decoding (PSDU) is performed using a subsampling rate of 75 %.

Related Publications:
Low-Power Features for H.264/AVC Hardware Encoders

N-M. Nguyen, E. Beigne, S. Lesecq, D-H. Bui(VNU), N.-K. Dang(VNU), X-T. Tran(VNU)

ABSTRACT: The H.264/AVC standard has been successfully used in a wide range of applications. In hardware design for H.264/AVC video encoders, power reduction is currently a tremendous challenge. This paper presents a survey of different H.264/AVC hardware encoders focusing on power features and power reduction techniques to be applied. A new H.264/AVC hardware encoder, named VENGME, is proposed. This low power encoder is a four-stage architecture with memory access reduction, in which, each module has been optimized. The actual total power consumption, estimated at RTL level, is 19.1mW.

Because of significant bit rate reduction in comparison to the previous video compression standards, the H.264/AVC has been successfully used in a wide range of applications. In hardware design for H.264/AVC video encoders, power reduction is currently a tremendous challenge. We present a survey of different H.264/AVC hardware encoders focusing on power features and power reduction techniques [1]. Table 1 compares several state-of-the-art solutions. Various features are presented but focus is on power features. Profile and resolution obviously influence the operating frequency and so the power consumption. Indeed, encoders that support multiple profiles [Lin08A1] or multiple resolutions [Chen09A, Lin08A1, Moch07A, Chan09A, Kim11P] operate at different frequencies and exhibit different power consumptions. When comparing power figures, the resolution and profile that the encoders support have to be taken into account. Low-power techniques [Chen09A, Moch07A] and strategies to reduce memory access [Kim11P, Lin08A1] show promising power consumption figures [Chen09A, Lin08A1, Moch07A]. Recent encoders with low-power features [Chen09A, Chan09A], with even smaller area cost, seem more suitable for mobile applications.

We also introduce a new H.264/AVC hardware encoder [2], named VENGME, see Figure 1. This low power encoder is a four-stage architecture with memory access reduction, in which each module has been optimized. The actual total power consumption, estimated at RTL level, is 19.1mW.

In the platform, some blocks, including the Entropy Coder - Network Abstraction Layer data packer (ECNAL) module, present nice power figures and performances similar to the state-of-art designs. However, its imbalance workload and different power control modes enable applying Dynamic Power Management (DPM) methods, for instance like Dynamic Voltage-Frequency Scaling (DVFS) or Dynamic Frequency Scaling (DFS).

![Figure 1: Architecture of the VENGME H.264 encoder.](image)

Table 1: Survey of H.264 Encoder Architectures

<table>
<thead>
<tr>
<th>Target</th>
<th>Lin08a</th>
<th>Chen08a</th>
<th>Chen09a</th>
<th>Chen06a</th>
<th>Chen07a</th>
<th>Lin08b</th>
<th>Chen08b</th>
<th>Kim11b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile</td>
<td>Baseline, level 4</td>
<td>High profile, SVC</td>
<td>High, level 4.1</td>
<td>Baseline, level up to 3.1</td>
<td>Baseline</td>
<td>Baseline, level 3.2</td>
<td>Baseline, high level</td>
<td>Baseline</td>
</tr>
<tr>
<td>Technology (nm)</td>
<td>UMC 180, 1P8M CMOS</td>
<td>UMC 90 1P9M</td>
<td>45</td>
<td>UMC 180, 1P8M CMOS</td>
<td>TSMC 180, 1P9M CMOS</td>
<td>Renesas RX, 2P17-V7Q ALP</td>
<td>UMC 130</td>
<td>28</td>
</tr>
<tr>
<td>Frequency (MHz)</td>
<td>250</td>
<td>120</td>
<td>162</td>
<td>81 for SD, 180 for HD</td>
<td>N/A</td>
<td>54 for SD, 144 for HD</td>
<td>7.2 for CIF, 145 for 1080p</td>
<td>10-12-18-28 for CIF, 72-108 for HD/720</td>
</tr>
<tr>
<td>Memory (bits)</td>
<td>108.3</td>
<td>81.7</td>
<td>230</td>
<td>34.72</td>
<td>56</td>
<td>56</td>
<td>22</td>
<td>13.3</td>
</tr>
<tr>
<td>Power consumption (mW)</td>
<td>1410</td>
<td>306 for high profile, 411 for SVC</td>
<td>256</td>
<td>581 for SD, 785 for HD</td>
<td>50.3 for CIF 2 ref, 9.8-13.9 for CIF 1 ref, 64.2 for 720p720TV</td>
<td>64 for 720p HD</td>
<td>5.74 for CIF baseline profile, 242 for 1080p high profile</td>
<td>7.25 for CIF, 122-184 for HD/720</td>
</tr>
</tbody>
</table>

Table 1: Survey of H.264 Encoder Architectures

Related Publications:
FIFO-level based Dynamic Power Management and its Application to a H.264 encoder

Research topics: H.264/AVC, low-power, power management, FIFO, buffer


ABSTRACT: We propose a new Dynamic Frequency Scaling implementation to manage the power consumption, based on the occupancy level of a FIFO. The approach proposed is simple and application independent. The PI controller was first tested in the MATLAB environment and then designed using VHDL. Simulations have validated the control method proposed. Synthesized in technology FDSOI 28nm, the control method presents only 10.8% of Silicon area overhead with 1.6% of power consumption reduction. This preliminary result is appealing because it has been obtained without voltage scaling that will improve the power gain even more.

We propose a new method to manage the power consumption by scaling the frequency (DFS method) according to the status of a FIFO link between components. The method is based on control theory. The control technique for power reduction is applied to the ECNAL module that contains the entropy coder (EC) communicating with the network abstraction layer data packet (NAL) via a FIFO in a H.264 encoder. The proposed method can be used in any FIFO/buffer based system where modules transfer data via FIFO links. For each link, a “producer” module writes data into the FIFO when it is not full. The “consumer” module reads data in the same order the data is written until the FIFO is empty.

The actual number of packets in the FIFO also depends on the clock frequency of the consumer, as packets are of different length depending on the data at hand. However, for simplicity, the FIFO output flow Q2 is supposed proportional to f_c(k):

\[ Q2(k) = b f_c(k) \]

where b is a positive constant. From several experiments, b has been estimated equal to 20ns.

The fact that Producer or Consumer has to wait for the FIFO availability leads to a waste of power consumption. Thus, the control objective is to adapt the consumer frequency to keep the FIFO half-full during normal operation. Of course, in the end, the FIFO has to be emptied.

For the FIFO link under study, the input data Q1 is not controlled. Therefore, in the system model, Q1 is considered as a disturbance. The FIFO transfer model, without disturbance, in the x-domain, is as follows:

\[ \frac{Q2(x)}{Q1(x)} = G(x) = \frac{x^2}{x^2 - 1} \]

A discrete-time Proportional-Integral (PI) controller is selected to reject the “disturbance” Q1, to ensure a closed-loop functioning without static error and to tune the closed-loop system time response. The controller is modeled as:

\[ f_c(k) = E(k) = K_p + K_i \frac{z}{z-1} \]

Therefore, the closed-loop transfer function is given by:

\[ \frac{FIFO(z)}{R(z)} = \frac{1}{z^2 - 2 + b(K_p + K_i) + 1 + K_p b} \]

where R(z) is the Reference for the FIFO level.

The poles z1, z2 determine the system dynamic characteristics in closed-loop. Their numerical values depend on K_p and K_i values. Once the control designer has chosen z1, z2, K_p and K_i are computed.

The PI controller proposed here was first tested in the MATLAB environment and then designed using VHDL. Simulations in the MATLAB environment and in ModelSim have validated the proposed control technique. Techniques to simplify the implementation are applied. Figure 3 presents the architecture of the controller where z1 = 0.75, z2 = 0.5.

Figure 3: Architecture of the controller when z1 = 0.75, z2 = 0.5.

Being synthesized in technology FD-SOI 28nm using RC synthesis tool from Cadence, the controller presents only 10.8% of Silicon area overhead with 1.6% of power consumption reduction. This preliminary result is appealing because it has been obtained without voltage scaling that will improve the power gain even more.

Related Publications:
In intrinsically, by applying a forward body biasing, UTBB FDSOI can reach a given operating frequency at a lower VDD than bulk. An additional design step is achieved in this work by (1) increasing the frequency at low VDD thanks to a specific selection and design of standard cells regarding power vs. frequency performance and (2) dynamically tracking the maximum frequency to cope with variations.

In the context of wide voltage range DSPs, an innovative system composed of an on-line programmable time-to-digital converter (TMFLT-R) correlated to off-line timing slack sensors (TMFLT-S) is proposed, at low area cost, and shows good functionality and accuracy even at low voltage down to 0.6V. 128 TMFLT-S, representing only 0.9% of the total number of flip-flops, are directly instrumenting flip-flops inside the core. However, during regular processing TMFLT-S paths may not be activated. To cope with, a less accurate, but always active, TMFLT-R sensor, based on a time-to-digital converter, is embedded and the FMAX control relies on its digital response (SIG) generated with a precision of 20ps. The overall TMFLT monitoring results on silicon, and after calibration, in an FMAX estimation error of +8/-6% at 0.6V and +4.1/-2.9% at 1V.

In the context of wide voltage range DSPs, this paper describes a 32-bit DSP fabricated in 28nm UTBB FDSOI technology. Body Biasing Voltage (VBB) scaling from 0V up to +/-2V (Pwell/Nwell) decreases the DSP core VDDMIN to 397mV and increases clock frequency by +400%@500mV and +114%@1.3V. In addition to technology gain, dedicated design features are included to increase frequency over the full VDD range in case of parameter variations.

Fast Pulse-triggered Flip-Flops (Pulsed-FFs) with small data-to-output (D-Q) delay, were designed to complete the optimized library. The smallest pulse width was defined with statistical simulations, to ensure robustness to 3σ variations down to 275mV. Pulsed-FFs silicon measurements were performed on a test scribe and showed correct functionality at 275mV and a measured D-Q delay of 1.3ns@0.3V. To maximize energy efficiency over the full voltage range in case of PVT variations, accurate and low cost FMAX estimation techniques are mandatory during circuit activity. We propose, in this work, two FMAX tracking solutions based on (1) replica path cloning (CODA) and (2) timing slack monitoring (TMFLT), both presenting a good accuracy at 1.0V and the second functional down to 0.6V. For both solutions, and over the full VDD range, the objective is to prevent timing failures and to reduce margins required for PVT variations during DSP operation. A Power-Variability Controller (CVP) is used to control, program and calibrate CODA and TMFLT and stores the captured timing information. For each frequency point, VDD and VBB are dynamically scaled to reach a minimum energy point.

In the context of wide voltage range DSPs, an innovative system composed of an on-line programmable time-to-digital converter (TMFLT-R) correlated to off-line timing slack sensors (TMFLT-S) is proposed, at low area cost, and shows good functionality and accuracy even at low voltage down to 0.6V. 128 TMFLT-S, representing only 0.9% of the total number of flip-flops, are directly instrumenting flip-flops inside the core. However, during regular processing TMFLT-S paths may not be activated. To cope with, a less accurate, but always active, TMFLT-R sensor, based on a time-to-digital converter, is embedded and the FMAX control relies on its digital response (SIG) generated with a precision of 20ps. The overall TMFLT monitoring results on silicon, and after calibration, in an FMAX estimation error of +8/-6% at 0.6V and +4.1/-2.9% at 1V.

Figure above shows frequency DSP silicon results obtained by performing a 1024 FFT using the register array configuration. The frequency curve shows 2.6GHz@1.3V(VDD)@2V(VBB) and 460MHz@397mV. The use of an efficient VBB (boost) above 1.5V decreases VDDMIN from 479mV down to 397mV. Energy per operation illustrates that, for a fixed energy budget at 100pJ/cycle, we can increase the frequency by 59% and reduce the energy by 17% to 20% for a fixed frequency. We obtain the highest ever reported operating frequency at 0.4V, close to the low peak energy efficiency measured at 62pJ/op at 0.46V.

Related Publications:
Power management through DVFS and dynamic Body Biasing in UTBB FDSOI circuits

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

ABSTRACT: The emerging SOI technologies provide an increased body bias range compared to traditional bulk technologies, opening new opportunities. Thus, from the power management perspective, a new degree of freedom is added to the supply voltage and clock frequency variation, increasing the complexity of the power optimization problem. A method has been proposed to manage the power consumed in an FD-SOI circuit through supply and body bias voltages, and clock frequency variation.

During the last years, several works have proven the efficiency of using Body Bias in UTBB FDSOI to reduce power consumption in integrated digital circuits. This work evaluates the combination of Dynamic Voltage and Frequency Scaling (DVFS) with Dynamic Body Bias (DBB) to manage power consumption at run-time. In DVFS approaches, supply voltage (Vdd) and clock frequency (F) are dynamically adjusted through on-chip actuators, while through DBB the body bias voltage (Vbb) is regulated to adjust the threshold voltage. However, most of actuator used in DVFS and DBB are discrete, i.e. they are able to reach a limited number of output values (F, Vdd, Vbb).

This work analyses the impact of the discretization of actuators and proposes a method to manage a system where at least one actuator is able to apply only a few set of values [1]. The first step of the proposed method consists on selecting the best (F, Vdd, Vbb) configurations belonging to a Piece-Wise Convex Subset (PWCS) in the power consumption profile in function of clock frequency (P(F)). During the circuit execution, only (F,Vdd,Vbb) configurations belonging to the identified PWCS are used, the convexity property of this subset ensuring optimality in terms of power consumption. Additionally, if some target frequency is not reachable by configurations in the PWCS, a dithering or hopping execution is used between two configurations in the PWCS, one at a higher frequency and another at lower frequency.

The proposed power management method has been successfully used in a circuit combining DVFS and DBB with a set of 4 reachable supply voltages and a wide-range body-bias (from 0V up to 1.5V) [2]. Results show gains reaching 18%, obtained out of the PWCS. These experiences prove that body biasing can overcome the limitations of using only few supply voltage values in the DVFS. Moreover, these results prove that the whole body bias range is not required at every voltage level to reach optimality and to cover a wide frequency range as shown in Figure 1.

Figure 1: Total power consumption, supply voltage and selected body bias ranges for a DSP in 28nm UTBB FDSOI [2].

Related Publications:
Early Design Stage Thermal Evaluation and Mitigation: the Locomotiv Architectural Case

Research topics: Thermal aware design flow

T. Sassolas, C. Sandionigi, A. Guerre, A. Aminot, P. Vivet, H. Boussetta, L. Ferro, N. Peitier

In order to offer more computing power to modern SoCs, transistors keep scaling down with new technology nodes. Consequently, the power density is increasing, leading to higher thermal risks. Thermal issues need to be addressed as early as possible in the design flow, when the optimization opportunities are the highest. In this work, we demonstrate that high-level architectural models, in conjunction with efficient power and thermal simulation tools, provide an adapted environment to analyze thermal issues and design software thermal mitigation solutions in the case of the Locomotiv MPSoC architecture.

Transistor size reduction induces greater power density resulting in higher chip temperature issues. Not considering heat dissipation devices, the common way to deal with thermal hotspots consists in evenly distributing temperature across the chip by managing the active portions of the circuit. Thus, the usage of the system will dictate its power and temperature evolution. Consequently, the architecture and software designs have to take into account these constraints as early as possible in the design flow. In order to efficiently deal with thermal issue we need an Electronic System Level (ESL) design environment that can take into account the system’s functionality, its power and thermal behaviors while modelling their mutual influences. In this paper, we present how we traded functional behavior and power accuracy to offer a fast environment to analyze thermal issues and design software thermal mitigation solutions for the Locomotiv MPSoC architecture [1]. The environment is composed of a Programmer’s View Loosely-Timed (PVLT) model tightly coupled with “Aceplorer”, a commercial ESL power and thermal analysis and optimization tool, and ”AceThermalModeller”, a compact thermal model generation tool, both developed by DOCEA Power.

An "Aceplorer" power model is composed of one or more power states for each component. For every power state, the user provides an analytical model for both leakage and dynamic consumptions. A particular effort was put into modelling the various power modes of the processing units, including a Vdd-Hopping mode dependent on the frequency state. RTL simulations were used to populate the analytical model. The impact of the temperature on both leakage and dynamic currents was modelled with respectively an exponential and linear regression from the power characterization data obtained for different temperature corner cases.

To model the thermal behavior of the system, a physical description of the geometry using rectangular cuboids and detailing every composing material with its thermal properties was made in "AceThermalModeler". This description was automatically processed to obtain a Dynamic Compact Thermal Model (DCTM). With such a DCTM the evaluation of the temperature is greatly accelerated while keeping a sufficient level of accuracy for ESL thermal evaluation. The DCTM can be imported in "Aceplorer" to simulate temperature effects. The Locomotiv chip is packaged into an SBGA304 from Amkor. Its geometrical description was conducted along with JEDEC standardized dimensions. The thermal properties of chemical elements were taken from literature while for compound elements, such as protection glue, they were extrapolated from reseller datasheet. The die floorplan is shown in fig.1.

Our PVLT simulator, based on an x86 implementation of the HARS runtime [2], compiles into a single executable the Locomotiv application, the runtime software, the runtime Hardware Abstraction Layer (HAL) code and the OSCI SystemC library. Application and runtime codes are executed on the x86 host machine. The x86 cycles are monitored during the execution using the RDTSC instruction. The target STxP70 cycles are extrapolated using a rule-of-thumb based on known processors Instructions Per Cycles (IPC). An emulated Locomotiv architecture using a Zebu-Server from Synopsys was used to validate the relative accuracy of our model.

![Figure 1: Power and temperature of the Locomotiv Chip executing 4 pedestrian detection frames (left) and temperature repartition on the die floorplan (right).](image)

Relaxing the accuracy on both the behavior modelling and the power accuracy to gain in simulation speed and deliver an efficient development environment for thermal mitigation. For a complete pedestrian detection application running on the Locomotiv architecture we were able to compare 2 thermal mitigation schemes applied during 10 simulated minutes for only twice that simulation time. We also showed that trivial thermal management is not sufficient for time-constrained application. So insight on the system functional, power and temperature behaviors is compulsory at the electronic system level. In the future, we plan to extend this co-simulation environment with co-emulation to bring both accuracy and speed for thermal mitigation development later in the flow. We also plan on studying the benefit of our environment for ageing evaluation.

Thermal Modeling and Exploration of 3D Circuits: Application to a Memory-on-Logic 65nm TSV middle circuit

C. Santos, P. Vivet, P. Colonna, N. Peltier (DOCEA), P. Coudrain (ST), R. Reis (UFRGS), L. Benini (Unibo)

Technology scaling implies power consumption densification, resulting in increased thermal effects. This is even worse for 3D architectures where power densities are larger and thermal dissipation is reduced with thin circuit substrates including TSVs. We present a thermal modeling methodology, based on ATM tool, offering early and accurate thermal system analysis; and an exhaustive thermal exploration of the 3D technology impact on die temperature, such as die thinning, TSV array induced hotspots, mitigation using Heat Spreaders. The study is carried on a Memory-on-Logic 3D circuit, including comparison of simulations and silicon measurements.

Modern electronic systems require more and more computing power which is achieved by technology scaling and architectural evolution thanks to Many-Core architectures. However the technology scaling comes with a densification of the power consumption, resulting in increased thermal dissipation. Current system presents also dynamic behavior in the power dissipation profiles, limiting systems’ thermal predictability. Thermal mitigation solution must be developed in order to control both power and thermal dynamic profiles. This is even more the case with 3D architectures where stacking multiple dies further increase power density within the 3D stack.

Using a Memory-on-Logic 3D circuit [5], a fast and accurate thermal modeling design flow has been developed, and a systematic thermal exploration of the impact 3D parameters has been carried out and compared to silicon results.

An exhaustive exploration of the 3D technology impact on die temperature has been performed [2], including analysis of: power density, die thinning, die underfill coupling, and with comparison of 3D die versus 2D die. Thermal mitigation of thermal hotspots has been proposed by using new materials like Graphite based Heat Spreader. This ultra thin anisotrop material increase the horizontal heat transfer leading to reduced thermal hotspots. Proposal has been validated in measurements and simulations (fig.2).

Thermal behavior analysis must be performed at system level and before design sign-off. A thermal modeling flow has been developed using ATM tool [1]. The methodology is based on accurate and compacted physical description of the full system, from large scale components such as board and package to fine grain structures like TSVs and µ-bumps. The thermal analysis has been applied to WIOING (fig.1) [5], a WideIO compatible 65nm Memory-on-Logic 3D circuit, and validated against silicon results. The circuit contains four memory controllers in the center, one per WideIO memory channel, plus the corresponding TSV and µ-bump matrices to connect to a WideIO compatible DRAM memory.

In 3D architectures, it has been often proposed to use TSVs in order to reduce thermal hot spots. When TSVs are properly modeled with their SiO2 layer, TSVs induce actually a decreased horizontal heat transfer coefficient for a light increase of the vertical heat path. The TSV will finally lead to larger thermal hotspots in case of lateral blockage, contrarily to the common believe [3]. Measurements on a dedicated testchip and systematic simulations (fig.3) have been carried out leading to the same conclusions.

Finally, using the same Memory-on-Logic 3D circuit, a thermal model has been developed allowing predicting system behavior [4]. The proposed identification technique could be integrated in the embedded software for thermal prediction and mitigation.

Related Publications:
3D stacked integrated circuits based on Through Silicon Vias (TSV) are promising with their high performances and small form factor. However, these circuits present many testability issues, such as test access mechanism & test scheduling, and test pattern generation in the overall 3D circuit fabrication steps. Various test standards from the industry co-exist today, including a novel standard currently under definition targeting 3D testability. For 3D test, we propose to use IEEE P1187 (IJTAG) test standard, allowing easy test pattern retargeting from 2D (pre-bond test) to 3D (post-bond test), while being compatible with on current test standards.

The proposed 3D DFT architecture is based on the IEEE 1687 standard (fig.2) : two different test access mechanism are provided : IJTAG for testing on chip instruments such as Boundary Scan interfaces for test of 3D interfaces, and test of on-chip elements such as memory BISTs, and regular full scan paths for logic test. The test architecture integrates TAP controllers and JTAG ports. The switch between the pre-bond and post-bond test modes is ensured by the mean of multiplexers that select test signals from pads for pre-bond level to 3D connections for post-bond level [2][3].

For 3D test, the first requirement is to define standardized test interfaces to allow compatible test access mechanisms between the different dies in a 3D stack circuit, and facilitate exchange of test information between Design-for-Test engineers, Test CAD tools, and Test engineers. An IEEE 3D-Test Working Group has been created in 2010, called IEEE 1838, to specify the test interfaces for 3D circuits. For 3D test, we propose to use the IEEE P1687 test, also called IJTAG (fig 1). The association of IEEE P1687, based on existing IEEE 1149.1 supported by advanced CAD tools, and of IEEE 1838 (the on-going standardization of 3D test interfaces) will not only guarantee the required 3D test compatibility but also will bring two additional advantages: i) easy test pattern retargeting from 2D (pre-bond test) to 3D (for post-bond test) and ii) enhanced flexibility in test concurrency thanks to the dynamic selection of instruments through IEEE P1687 circuitry [3].

Related Publications:
Shadow-scan design with low latency overhead and in-situ slack-time monitoring

Research topics: Scan design, In-situ slack-time monitoring, Timing violations

S. Sarrazin, S. Evain, I. M.-Panades, A. Valentian, S. Pajaniradja, L. A. de B. Naviner, V. Gherman

ABSTRACT: Shadow-scan solutions are proposed in order to facilitate the implementation of faster scan flip-flops (FFs) with optional support for in-situ slack-time monitoring. These solutions can be applied to system FFs placed at the end of timing-critical paths while standard-scan cells are deployed in the rest of the system. Automated scan stitching and test pattern generation can be performed transparently with commercial tools. The generated test patterns cover not only the mission logic but also the monitoring infrastructure. The latency of ic’t99 benchmarks was reduced with up to 10% while the stuck-at fault coverage was preserved.

Slack-time reduction is a way to improve the power consumption of synchronous sequential circuits without affecting their performance. In the presence of circuit wear-out, supply voltage fluctuations and temperature variations, aggressive slack-time reduction can be achieved based on adaptive voltage and frequency scaling with feedback from in-situ slack-time monitoring. Moreover, shadow-scan solutions for in-situ slack-time monitoring can also be used to reduce the latency of scan FFs [1].

A new shadow-scan solution is proposed, as illustrated in Fig. 1, which (a) facilitates the implementation of faster scan FFs, (b) enables in-situ slack-time monitoring and (c) can be transparently handled by commercial tools for automated scan stitching and automated test pattern generation (ATPG). This solution relies on a shadow sequential element (SSE) associated to a system FF. In scan mode, test data is first transferred to the SSE and subsequently to the system FF with the help of local set and reset operations controlled by the associated SSE. In the case of a system FF with functional reset, a local reset operation is triggered by a logic 0 value scanned in its SSE. Otherwise, a local set operation is forced via the scan-enable signal.

There is no need to impose minimum delay constraints or to provide delayed clock signals. Compared to [1], it is not necessary to apply a reset pulse to transfer scan data from the SSE to the system FF. System FFs with standard-scan design can still be used if they are not placed at the end of timing-critical paths.

A shadow-scan solution without in-situ slack-time monitoring capability is also proposed, as shown in Fig. 2, which enables faster implementations for a majority of ic’t99 benchmarks.

In the presence of in-situ slack-time monitoring support, a combination of both solutions allowed to improve the latency of ic’t99 benchmarks with up to 10% as compared to circuit versions with full standard-scan design. In the absence of in-situ slack-time monitoring support, the circuit latency could be improved with up to 6%. The proposed solutions do not hinder the use of commercial tools for automated scan stitching and ATPG. The generated test patterns could cover the faults in the mission logic and the monitoring infrastructure. The achieved stuck-at fault coverage was the same as in the case of full standard-scan design. Limited variations in the test sequence length were observed when support for in-situ slack-time monitoring was provided.

Related Publications:
Flip-Flop Selection for In-Situ Slack-Time Monitoring based on the Activation Probability of Timing-Critical Paths

Research topics: Dynamic variations, On-line monitoring, In-situ slack-time monitoring

S. Sarrazin, S. Evain, I. Miro-Panades, L. Alves de Barros Naviner (IMT), V. Gherman

ABSTRACT: In-situ slack-time monitoring is used to enable ambitious power management under circuit wear-out and dynamic variations. Given a limited hardware budget, it becomes crucial to be able to select the most appropriate places for in-situ slack-time monitoring. Two metrics are proposed to guide the selection of a set of flip-flops (FF) with monitors. The goal of these metrics is to maximize the ratio of clock cycles with at least one monitor activated and the number of activated monitors per clock cycle. The monitor activation probability is evaluated with the help of timing simulations as the signal propagation probability along monitored timing-critical paths.

Dynamic variations represent a serious challenge to the design of performant and energy efficient synchronous circuits. One can distinguish fast changing variations which may be triggered by high frequency supply voltage droops and slow changing variations e.g. variations induced by circuit wear-out or lower frequency temperature and/or supply voltage variations. The conventional approach to deal with such variations is to insert a temporal guardband or an increased slack-time in the clock period. Unfortunately, in the absence of other mitigation techniques, a significant slack-time may be required, e.g. up to 20% of the clock period, with an upward trend led by the continuous technology scaling.

Slack-time minimization became synonym for supply voltage reduction and can be used to moderate the dissipated power and slowdown the circuit wear-out. Adaptive voltage and frequency scaling with feedback from in-situ slack-time monitoring may be used to cope with relatively slow changing variations. A natural solution is to place in-situ slack-time monitors close to all sequential elements with incoming timing-critical paths or susceptible to become timing-critical due to wear-out or manufacturing variability. In latency-constrained circuits with large ratios of timing-critical paths, this methodology may result in large area overheads and minor power improvements.

Candidate FFs for in-situ slack-time monitoring are evaluated according to the probability that signal transitions are received along timing-critical paths. Such a signal transition may activate an in-situ slack-time monitor assigned to the destination FF and trigger a warning. Only slack-time monitors with error prediction are considered since they do not impose special error-recovery features or minimum path delay constraints. Such monitors provide a detection window during which late transitions of the monitored signal can be sensed. A large detection window provides a more robust design but may affect the logic and power overhead.

The monitor activation probability is estimated based on timing simulations with random stimuli applied to the primary circuit inputs and appropriately chosen clock periods in order to model slack-time degradation. When available, typical workload stimuli can also be applied.

Two metrics are proposed to evaluate the monitoring quality delivered by a set of FFs with slack-time monitoring: the expected ratio of clock cycles with at least one monitor activated (ERCMA) and the expected number of activated monitors per clock cycle (ENAMC). The ERCMA and ENAMC, respectively, can be used to estimate the temporal and spatial coverage of an in-situ slack-time monitoring scheme.

Based on these metrics, it is shown that the monitoring quality can be significantly improved if the size of the detection window of each in-situ slack-time monitor is correlated to the slack-time of the monitored timing-critical paths without affecting the circuit latency. The proposed methodology can be applied to different types of circuit nodes with in-situ slack-time monitoring and combined with other metrics in order to provide a certain (a) uniformity of the spatial distribution and (b) diversity of the monitored circuitry.

The impact of using correlated detection windows was simulated on a relatively large itc’99 benchmark circuit (b22) with a total of 1192 system FFs. As shown in Table I, the ENAMC and ERCMA metrics were boosted by a factor of 6 and 15, respectively, when late transitions (slack) as large as 2% of the most timing-critical path occurred within the monitoring detection window.

<table>
<thead>
<tr>
<th>Slack window</th>
<th>Detection window</th>
<th>Ratio of FFs with activated monitors</th>
<th>ENAMC</th>
<th>Ratio of simulated clock cycles with activated monitors</th>
<th>ERCMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2% constant correlated</td>
<td>1.00%</td>
<td>0.03</td>
<td>2%</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>-4% constant correlated</td>
<td>4.45%</td>
<td>1.82</td>
<td>33%</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>-4% correlated</td>
<td>4.45%</td>
<td>3.72</td>
<td>41%</td>
<td>0.40</td>
<td></td>
</tr>
</tbody>
</table>

Related publications:
Content-addressable memories (CAMs) are special memories with high-speed parallel search capability. Ternary CAMs (TCAMs) enable the storage and comparison of don’t cares besides 0s and 1s, property which can be used to reduce storage requirements. TCAMs are especially used in network routers to provide next-hop information or packet classification for flexible quality-of-service policies. 

Static random-access memory (SRAM) is still the prevailing storage infrastructure in TCAMs despite the existence of other technologies. Unfortunately, the SRAM cells are vulnerable to soft-errors. During a lookup operation in a TCAM, the presence of soft-errors may induce a false-hit, i.e. an incorrect address is returned, or a false-miss, i.e. no address location is returned despite the fact that a word compatible with the search word was stored in the TCAM.

In this work, a solution is proposed to improve the ECC-based protection of SRAM-based TCAMs against soft-errors. This solution is inspired by the following observations:

(a) A false-miss is difficult to detect since any word stored in the TCAM could correspond to a corrupted version of the search word and needs to be checked.

(b) A false-hit can be overcome when it returns the address of a matching word which is affected by errors that are correctable, e.g. with the help of an ECC.

(c) The probability that a soft-error induces a false-hit that points to an error-free matching word or a false-miss can be reduced if don’t care values are stored in a robust manner while 1 and 0 values can only be corrupted into a don’t care value.

The last requirement can be achieved with the help of asymmetric SRAM cells that have one of the states reinforced against soft-errors. For example, such solutions were proposed to improve the reliability of applications in which one of the two states of the SRAM cells occurs more often. A possible encoding of the values stored in a TCAM cell is presented in Fig. 1. If the 0 state of the SRAM cells is reinforced then (a) don’t care values will become robust to soft-errors, and (b) 1 and 0 values will have a low probability to be corrupted into any other value than a don’t care.

The usage of asymmetric storage-cells combined with an appropriate ternary value encoding, as illustrated in Fig. 1, are able to reduce the occurrence probability of false-misses or false-hits that point to an error-free matching word. For example, consider the word "-110001" stored in an ML where the symbol ‘-' stands for don’t care. Due to the storage-cell asymmetry, a soft-error can only transform a ‘1’ or a ‘0’ into a ‘-', e.g. the corrupted word may become "-101001" which still matches any search word, e.g. "1110001", that matches the error-free word. Lookup operations that involve such search words will not be affected by the considered soft-error. Only a search word, e.g. “1010001”, which does not match the error-free word but the corrupted word may induce a faulty behavior and this only in the absence of other matching word with higher priority. In the latter case, the outcome would be a false-hit that returns the address of the corrupted word.

Such false-hits can be mitigated if (a) the word stored in each ML is encoded with the help of an ECC and (b) the word at the address returned by a lookup operation is retrieved and verified as in the case of conventional memories protected by ECCs. Periodic scrubbing can also be performed in order to (a) avoid error accumulation in rarely searched TCAM words or (b) correct the rare, if any, soft-errors which could induce false-misses i.e. bit-flips of reinforced states.

**Figure 1 :** Schematic view of a match line (ML) in a NOR-type TCAM

---

**ABSTRACT:** Content-addressable memories (CAMs) enable the comparison of their entire content to a search word in one single access. Ternary CAMs (TCAMs) provide the possibility to store and handle not only 0’s and 1’s but also "don’t care" states. We propose a way to protect TCAMs implemented with static random-access memory (SRAM) cells against soft-errors. Asymmetric SRAM cells are used to reduce the probability that soft-errors affect "don’t care" states and corrupt 0s and 1s into anything else than a "don’t care."

**Research topics:** Ternary content-addressable memories, Soft-errors, ECC

---

**Related publications:**
4

Multi-Cores & Many-Cores Architectures & Software

Parallel Architectures
Optical Network on Chip
Task mapping & Scheduling
The complexity of SystemC virtual prototyping is continuously increasing. Accelerating RTL/TLM SystemC simulations is essential to control future SoC development cost and time-to-market. In this work, we present RAVES, a highly-parallel special-purpose multicore architecture that achieves simulation performance more efficiently by parallel execution of light-weight user-level threads on many small cores. We present a design study based on the RAVES virtual prototype running a co-designed custom SystemC kernel. Our evaluation suggests that a 64-core RAVES processor can deliver up to 4.47x more simulation performance than a high-end x86 processor.

The SKA is used to execute the simulation manager and to interleave multiple SystemC simulations. The SKA reduces kernel and synchronizations overheads. The RAVES SystemC kernel is fully compliant with SystemC 2.2.0. It has been optimized to support the execution of SystemC processes. It supports the creation of thread contexts, process allocation, process preemption and migration. Only the evaluation phase is executed in parallel on multiple cores. One kernel thread per core, named worker, is used as a container to locally execute user-level threads through ucontext primitives.

We designed a virtual prototype of RAVES within the SESAM environment [2]. SESAM is a SystemC/TLM simulation framework that eases the architecture exploration of multicore platforms. Our light-weight user-level threads and co-designed custom SystemC kernel bring to RAVES a unique capacity to accelerate RTL and TLM simulations on many cores. Our evaluation suggests that a 64-core RAVES processor can deliver better simulation performance than a high-end x86 processor. It can reach up an acceleration of 4.7x compared to a user-level thread SystemC implementation running on an i7-2600K with small RISC cores. The results on the TilePro64 architecture show that only a dedicated architecture can compete with the i7. We believe that the underlying parallelization approach can be generalized to enable a performance oriented system based on high-performance cores.

Related Publications:
With an ever increasing need for integrated computing power parallelism, and in view of the impact of die size on system yield, we studied how general purpose parallel computing applications could be mapped on a multi-die platform made of several chiplets stacked on an interposer, following different technological and architectural options.

Starting from the intrinsic computing performance of a 2GHz quad-core chiplet, and the associate traffic between chiplets for a distributed L3 cache, we defined six different microarchitectures using various interposer technologies, and assessed their system performance, power and area:

- passive interposer with low-impedance transmission lines;
- CMOS interposer with synchronous point-to-point links;
- idem with quasi-delay-insensitive asynchronous links;
- idem with a synchronous pipelined 2D-mesh network;
- idem with an asynchronous 2D-mesh network;
- optical interposer with silicon photonic devices.

The metrics were computed as closed-form equations for three levels: single point-to-point link (2 chiplets), a small-scale architecture (3-6), and a large-scale architecture.

First on the application side, the different options lead to a very different system latency (communication set-up, message completion). Synchronous solutions have a huge latency (up to 100 cycles) compared to other solutions due to low-frequency pipelining on the interposer. Asynchronous solutions have a latency similar to DC buffering across chip, which is still twice as high as high-speed serial solutions.

In terms of area, point-to-point lines on CMOS interposer suffer from the quadratic explosion in the number of links compared to the NoCs. High-speed links, electrical or optical, have a much lower footprint, and challenge the NoC in terms of area, staying around 10% of chiplet size.

Yet, the footprint is also constrained in terms of critical height, i.e. either the 3D interface height reported to chiplet height, or the cumulative line height at interposer bisection reported to interposer height. As before, the CMOS point-to-point lines, synchronous or asynchronous, rapidly require the complete interposer height to connect all chiplets. As for the high-speed transmission lines, they suffer from the chiplet interface on the perimeter, due to pitch constraints.

In terms of power consumption at system level for communication (Fig. 1), i.e. the sum of idle power (leakage, dynamic power of clock distribution for CMOS, laser and thermal tuning for optics) and useful switching power, reported to total power including computing, the optical NoC starts with a high static cost compared to the other solutions, but as the number of chiplets increase, the electrical NoC is sized accordingly to avoid contention, at the cost of higher power consumption.

This analysis shows that the metal interposer is best for a single chip-to-chip connection, such as a processor to high-density memory connection as in the Hybrid Memory Cube, but is not relevant for larger-scale integration. In that case, the active interposer appears as a viable solution up to 10-12 dies, but then suffers from NoC area and latency. The optical interposer, even in a simple single-writer multiple-reader solution, offers better scalability, and allows for denser integration at a similar power cost, for multiprocessors of 16 chiplets or more.

Figure 1: Influence of the interposer technology and architecture on the overall applicative system power (idle+dynamic power)

Related Publications:
The ongoing trend of integrating more and more processors on a single dice is present in all performance classes ranging from mobile applications to high performance computers (HPC). Especially HPCs require this technique for future performance scaling. Nevertheless, increasing the number of processors in one dice or one package does not solve all problems or might even introduce new ones.

In modern HPC projects the power consumption is a major factor not only for the computers architecture but for the funding of the project and construction side choice. Counter intuitively the biggest part of the power consumption is not due to the data processing but due to the data transport inside HPCs. For example a floating point operation needs about 0.1 to 0.05 pj/b while the on card data transport consumes already about 2-10 pj/b. Another bottleneck for future scaling of HPCs is the needed data rate for the communication between processors and memories.

Silicon photonics addresses both these problems and will be integrated into HPCs in the near future. Especially optical networks on chips (ONoCs) will have a big impact on the HPC market. There are two options when implementing ONoCs. The electronic and optical components are either combined on the same substrate (only favorable if an older technology node is sufficient for the computing tasks), or split into an optical interposer with flip-chipped electronic dice on it. An artistic representation of the approach followed by our internal HUBEO project is shown in Fig. 1.

Although electro-optical receivers have been developed for a long time, this new environment sets new demands and asks for new solutions. As the receiver will be only a subsystem of a large computing system it has to have a small footprint, small power consumption and high robustness. On the other hand the parasitic input capacitances (a well known bottleneck) are very small.

In order to optimize future receiver design, the optimal bitrate and optimal topology of TIAs has been studied in [1]. The general result of these studies is that there is a technology dependent speed optimum, which is below 20 Gbps for the studied 65nm CMOS technology node and that the feedbacked push-pull inverter topology seems to be favorable in this environment.

The already mentioned demands on the receiver are extended when the ONoC incorporates multiple writer single reader links. This means that dependent on the writing entity the strength and offset of the optical signal varies. In [2] a receiver front-end is presented, optimized for this problem. It is argued that the sizing ratio of the transistors inside the push-pull inverter has to be system dependent and that a transistor should be used as feedback device. This allows to widen the accepted signal range, by the control of the feedback transistors gate voltage, see Fig. 2.

Related Publications:
ABSTRACT: In the context of networks of massively parallel execution models, optimizing the locality if inter-process communication is a major performance issue. We propose two heuristics to solve a dataflow process network mapping problem, where a network of communicating tasks is placed into a set of processors with limited resource capacities, while minimizing the overall communication bandwidth between processors. Those approaches are designed to tackle instances of over one hundred thousand tasks in acceptable time.

With the end of the frequency version of Moore’s law, a new generation of massively multi-core microprocessors is emerging. This has triggered a regain of interest for the so-called dataflow programming models in which one expresses computation-intensive applications as networks of concurrent processes interacting through unidirectional FIFO channels. In this work [1], we focus on the problem of mapping a dataflow process network (DPN) on a clustered parallel microprocessor architecture composed of a number of nodes, each of these nodes being a small SMP, interconnected by an asynchronous packet network. A DPN is modeled by a graph where the vertices are tasks to place, and the edges represent communication channels between tasks. Vertices are weighted with one or more quantities which correspond to processor resources consumption and the edges are weighted with an inter-task communication outflow. The aim of our problem is to maximize inter-task communications inside SMPs while minimizing inter-node communication under capacity constraints to be respected in terms of task resource occupation on the SMPs. We present two methods: Greedy Task Wise Placement (GTWP) and Subgraphs placement. Each of those methods is able to tackle large instances of that problem in a reasonable amount of time.

In the Greedy Task Wise Placement all tasks are assigned one after another with regards to a notion of distance affinity. It is a one phase process. The Subgraphs Placement method is a two phase method: instead of assigning tasks one by one like the previous method, we generate a subgraph of connected tasks which are then placed on a node. The connected obtained subgraph is assigned to a node depending on the affinity between the subgraph and the nodes.

Two kinds of task graph topologies were used. First, a set of grid shaped task topologies, which correspond to dataflow computational networks like matrix products. Second, a task graph is generated out of logic gate networks resulting in the design of microprocessors. These configurations typically can be found in real life complex dataflow applications. The node layout is a square torus, hence the number of nodes in all instances is in a square dimension.

For graphs generated out of logic gate networks, the edge weights are the number of arcs between the corresponding elements in the original multigraph. For each pair of nodes (n, n’), the distance is the Manhattan distance between nodes n and n’. In those experimentations, all instances are limited to one resource and the resource occupation of every tasks in arbitrarily set to 1. It has been developed at CEA List for the ZC toolbox for the MPPA manycore platform.

We compare our methods with that of [1]. We denote this method as Partition and Place (P&P).

The results are displayed in the following table:

<table>
<thead>
<tr>
<th>Graphs</th>
<th>Number of vertices</th>
<th>Mapping Cost GTWP</th>
<th>Time</th>
<th>Mapping Cost Subgraphs</th>
<th>Time</th>
<th>Mapping Cost P&amp;P</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grids</td>
<td>144</td>
<td>41</td>
<td>2.3ms</td>
<td>75</td>
<td>382μs</td>
<td>37</td>
<td>0.02s</td>
</tr>
<tr>
<td>Logic Gate Networks</td>
<td>21,183</td>
<td>2,100,000</td>
<td>0.17%</td>
<td>2,540,000</td>
<td>1.3ms</td>
<td>2,500,000</td>
<td>2s</td>
</tr>
<tr>
<td>Networks</td>
<td>10,000</td>
<td>16,000</td>
<td>3%</td>
<td>18,793</td>
<td>0.33%</td>
<td>45,613</td>
<td>240s</td>
</tr>
<tr>
<td>Logic Gate Networks</td>
<td>1,065,000</td>
<td>1,058,000</td>
<td>9.66%</td>
<td>2,205,000</td>
<td>48.66%</td>
<td>1,250,000</td>
<td>4.85%</td>
</tr>
<tr>
<td>Networks</td>
<td>231,246</td>
<td>NONE</td>
<td>NONE</td>
<td>5,113,634</td>
<td>416,568</td>
<td>NONE</td>
<td>NONE</td>
</tr>
</tbody>
</table>

We can notice that the higher the number of tasks, the more the solution quality of GTWP increases compared to the solution quality of the P&P. There is, depending on the instance, a relative speedup of 67.

For the Subgraphs placement method, the runtimes are several orders of magnitude faster than P&P, while providing solutions whose quality tends to get comparatively similar or better on the largest instances. GTWP provides better results than the Subgraphs method, while the latter runs faster and scales easily on very large instances.

The increase in terms of compared solution quality between our methods and P&P finds its explanation in two different aspects. First, as the partitioning phase of P&P does not take node distance into account, tasks are gathered together with no knowledge of the location of the processor in the topology. Thus, choices made during this phase may undermine the overall solution quality. In the opposite, the distance affinity notion we use in GTWP allows us to take profit of the topology and avoid many bad choices. Second, even not taking profit from the node distances, the Subgraphs placement method has the advantage that it tries to avoid placing singletons or very small sub graphs, while the last 10% (or perhaps more) of the tasks to be assigned in P&P may probably not be efficiently assigned, leading to a drop in quality.

Related Publications:
There is an increasing interest in developing applications on multiprocessor platforms due to their broad availability and the looming horizon of many-core chip, such as the MPPA-256 chip from Kalray (256 cores) or the SThorm chip from STMicroelectronics (64 cores). Given the scale of these new massively parallel systems, programming languages based on the data-flow model of computation have strong assets in the race for productivity and scalability. Nonetheless, as streaming applications must ensure data-dependency constraints, scheduling has serious impact on performance. Hence, multiprocessor scheduling for data-flow languages has been an active area and therefore many scheduling and resource management solutions were suggested.

We introduce two classes of STP schedules based on two different granularities. The first schedule, denoted as STP$_{qi}$, are based on the repetition vector $q_i$ without including the sub-tasks of actors. A remaining schedule, denoted as STP$_{ri}$, have a finer granularity by including the sub-tasks of actors. It is based on the repetition vector $r_i$.

The effect of Self-Timed Periodic (STP) scheduling can be modeled by replacing the period of the actor in each level by its worst-case execution time under periodic scheduling. The worst-case execution time is the total time of computation and communication parts of each actor. For a graph $G$, a period $\Phi$, which represents the period, measured in time-units, of the levels in $G$, is given by the solution to:

$$\Phi \geq \max_{j=1}^{\alpha} (W_j + \phi_j)$$

where $\alpha$ is the number of levels, $W_j$ is the maximum workload and $\phi_j$ is the worst-case communication time of all levels in the Timed Graph $G$.

Let $a_1$ denote the level-1 actor; $a_1$ will complete one iteration when it fires $q_1$ times. Assume that $a_1$ starts executing at time $t = 0$. Then, by time $t = \Phi \geq q_1 \omega_1$, $a_1$ is guaranteed to finish one iteration in a self-timed mode (start the next sub-task immediately after the end of the precedent); $a_1$ will also generate enough data such that every actor $a_k \in V_2$ can execute $a_k$ times (i.e. one iteration). By repeating this over all the $\alpha$ levels, a schedule $S_\alpha$ shown in Figure 1 is constructed:

![Figure 1: Schedule $S_\alpha$ constructed by aapplication of the algorithm.](image)

We evaluate the proposed STP representation using a set of 10 real-life applications. Table I shows the latency obtained under STS, SPS, STP$_{qi}$, and STP$_{ri}$ schedules as well as the improvement of these policies compared to the SPS model. For the STP$_{qi}$ approach, we have an average improvement of 35.8% compared to the SPS model for all the applications. For the STP$_{ri}$ approach, we have an average improvement of 35.8% compared to the SPS model for all the applications.

<table>
<thead>
<tr>
<th>Application</th>
<th>N</th>
<th>Q</th>
<th>max($q_i a_i$)</th>
<th>STS</th>
<th>SPS</th>
<th>STP$_{qi}$</th>
<th>STP$_{ri}$</th>
<th>$E_{eff_{STP}}$ (%)</th>
<th>$E_{eff_{STP}}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC1</td>
<td>12</td>
<td>1200</td>
<td>2500</td>
<td>7200</td>
<td>5400</td>
<td>27000</td>
<td>60.2</td>
<td>32000</td>
<td>30.8</td>
</tr>
<tr>
<td>FFT</td>
<td>4</td>
<td>900</td>
<td>23000</td>
<td>36000</td>
<td>2300</td>
<td>11.5</td>
<td>30000</td>
<td>25.9</td>
<td></td>
</tr>
<tr>
<td>Beamformer</td>
<td>4</td>
<td>7800</td>
<td>9500</td>
<td>25200</td>
<td>23400</td>
<td>11.5</td>
<td>30000</td>
<td>30.6</td>
<td></td>
</tr>
<tr>
<td>Pilotebank</td>
<td>17</td>
<td>6000</td>
<td>113430</td>
<td>124792</td>
<td>1245000</td>
<td>1247370</td>
<td>0.6</td>
<td>1472400</td>
<td>0.6</td>
</tr>
<tr>
<td>MPS</td>
<td>5</td>
<td>25320</td>
<td>9600</td>
<td>10800</td>
<td>141120</td>
<td>5760</td>
<td>33.3</td>
<td>108</td>
<td>85.3</td>
</tr>
<tr>
<td>Sample-rate</td>
<td>5</td>
<td>35320</td>
<td>3825000</td>
<td>644000</td>
<td>1558000</td>
<td>1525000</td>
<td>5.6</td>
<td>1525000</td>
<td>5.6</td>
</tr>
<tr>
<td>H.263 Dec</td>
<td>5</td>
<td>3376</td>
<td>40000</td>
<td>40000</td>
<td>40000</td>
<td>31.3</td>
<td>40000</td>
<td>31.3</td>
<td></td>
</tr>
<tr>
<td>Bipartite</td>
<td>4</td>
<td>144</td>
<td>252</td>
<td>293</td>
<td>576</td>
<td>504</td>
<td>25.3</td>
<td>504</td>
<td>25.4</td>
</tr>
<tr>
<td>Satellite</td>
<td>22</td>
<td>5280</td>
<td>1056</td>
<td>1314</td>
<td>58080</td>
<td>11616</td>
<td>81.9</td>
<td>11616</td>
<td>81.9</td>
</tr>
</tbody>
</table>

Table 1: Benchmarks used for evaluation

Related Publications:
Comparing the StreamIt and ΣC Languages for
Manycore Processors

Research topics: Many-core, Stream languages

XK. Do, S. Louise (CEA LIST), A. Cohen (INRIA)

ABSTRACT: Stream programming fits embedded manycore systems' requirements in terms of parallelism, functional determinism, and local data reuse. Statically or semi-dynamically scheduled stream languages like e.g. StreamIt and ΣC can generate very efficient parallel code, but have strict limitations with respect to the expression of dynamic computational tasks. This paper compares two state-of-the-art stream languages, StreamIt and ΣC, with the aim of better understanding their strengths and weaknesses, and finding a way to improve them. We also propose an automatic conversion method and tool to transform between these two languages.

The emergence of chip multiprocessor architectures integrating hundreds or thousands of processing units in a single chip pushes for profound changes in the expression of scalable parallelism up to millions of concurrent tasks and its automatic exploitation on complex, often distributed memory hierarchies. But programmers currently lack a clear path to navigate the landscape of many-core programming models and tools, and programming tools are unfortunately lagging behind the fast paced architectural evolutions.

StreamIt and ΣC provide an advanced compiler for streaming applications by performing stream-specific optimization, matching the performance of achievable by an expert. Nevertheless, one of the disadvantages of these languages is the difficulty to reconfigure the static data-flow graph. Such limitations have been analyzed in detail on a wireless network application where dynamic agents need to be inserted to tolerate high noise on an input channel, cleaning up the signal.

These restrictions led to the quest for a compromise between the language expressive power and the compiler's mapping capabilities. Our work complements these approaches, building on a systematic study of StreamIt and ΣC and on the latter's tool flow targeting a real-world many-core processor (Kairaay MPPA) to offer additional insights into the design of expressive streaming languages that remain suitable for compilation on a many-core processor. We also discuss on embedded system requirements regarding the expression of non-functional properties in these languages.

First, we can say that StreamIt and ΣC have a lot of similarities. The core of these languages is interconnected agents and filters, both are autonomous entity with its own non-shared address space with other tasks of the application. However, both languages support only static software architecture (links between tasks, data exchange amounts are determined at compile time). Other similar aspects between these languages are the real-time requirements which are not often well taken into account, although it can be enhanced quite easily. Besides these similarities, we can recognize a lot of differences between these two languages. While StreamIt tries to create a SDF graph of connected filters, the model of computation of ΣC is CSDF. As a compiler, ΣC on MPPA can be compared to the StreamIt/RAW compiler. However, the execution model supported by the target is different: dynamic tasks scheduling is allowed on MPPA; the communication topology is arbitrary and uses both a Networks on Chip (NoC) and shared memory.

After studying StreamIt and ΣC, a method and tool were developed to convert StreamIt benchmarks in ΣC, as can be seen in Listing 1 and Listing 2. This work aims to create a library of benchmarks for ΣC.

Figure 1: Process Network topology of a StreamIt and ΣC graph, as shown in Listing 1 and Listing 2. In this topology, multiple iDCT agents (iDCT) connected to one split (mSplit) and one join (mJoin).

Listing 1: Topology building code of the StreamIt graph

```c
int->int splitjoin iDCT4x4_1D_X_fast_fine() {
    split roundrobin(4);
    for (int i = 0; i < 4; i++) {
        add iDCT4x4_1D_row_fast();
        join roundrobin(4);
    }
}
```

Listing 2: Topology building code of the ΣC subgraph

```c
subgraph iDCT4x4_1D_X_fast_fine() {
    interface {
        in<int> input;
        out<int> output;
    }
    map {
        int i;
        agent mSplit= new Split<int>(4,2);
        agent mJoin= new Join<int>(4,2);
        for (i = 0; i < 4; i++) {
            agent iDCT= new iDCT4x4_1D_row_fast();
            connect(mySplit.output[i], iDCT.input);
            connect(IDCT.output, myJoin.input[i]);
            connect(input, mySplit.input);
            connect(myJoin.output, output);
        }
    }
}
```

Related Publications:

60
5

Dynamic & Real-time Software

Dynamic Compilation
Parametric Dataflow
Real Time Software
Micro-architectural simulation of embedded core heterogeneity

Research topics: Power, performance, micro-architecture, ARM, heterogeneous cores

Fernando A. Endo, Damien Couroussé and Henri-Pierre Charles

Energy consumption is the major factor limiting performance in embedded systems. In addition, in the next generations of ICs, heat or energy constraints will not allow to power all transistors simultaneously. Heterogeneous multicore systems represent a possible solution to this problem: the diversity of cores provides energy and performance trade-offs. While, core heterogeneity cannot be studied with simple one-IPC simulators, detailed (RTL) simulation is too complex and requires huge computing resources. This work is the first to develop a micro-architectural simulator capable of estimating energy and performance of in-order and out-of-order ARM cores.

In high-performance embedded systems, reduced energy consumption is essential for providing more performance. As current transistor technologies cannot efficiently reduce the power densities in ICs, next generations of processors will need to improve performance with only incremental increases of power budget.

A good way to increase the total throughput of a processor with a limited power budget is to replace big and power-hungry out-of-order cores by little and energy-efficient in-order ones. However, as single-thread performance is still important in the embedded market, core asymmetry is the solution to have both single-thread peak performance and acceptable power dissipation when the system load is high.

Studies predict that heterogeneous multicore or even manycore processors will be the only solution to improve energy efficiency and sustain the performance growth predicted by Moore’s law. To study such emerging systems, although instruction set simulators (ISSs) are fast and provide good estimations, they are not adapted, because they are mostly calibrated to a specific hardware. In contrast, detailed simulation at the RTL-level is quite accurate, but it lacks abstraction and requires huge computing resources. Finally in the midway, micro-architectural simulators offer a good trade-off between precision and simulation speed. With more abstraction than RTL-models, micro-architectural simulation is flexible enough to study varying pipeline configurations, which is not possible with ISSs or one-IPC simulators. Figure 1 illustrates trade-offs of simulation accuracy and speed.

Previous studies on core heterogeneity employed x86 and Alpha simulators. While x86 is representative of desktop- and server-class processors, Alpha processors were discontinued early in the 2000’s. The ARM ISA is more relevant in embedded computing ecosystems. The advantage of using ARM simulators is that researchers can evaluate their tools without porting them to other platforms.

To study dynamic code optimizations and auto-tuning in the context of growing embedded core heterogeneity, we developed a micro-architectural simulator of in-order and out-of-order cores. In terms of execution time estimation, our framework showed an average absolute error of 7 %, when simulating ARM Cortex-A8 and A9, running 10 and 20 configurations of modern and parallel benchmarks (PARSEC). Similar state-of-the-art simulators have average errors greater than 15 %.

Figure 2 shows the simulated energy and performance trade-offs of big.LITTLE CPUs (Cortex-A7 and A15), based on an Exynos 5 octa processor. For the Dhrystone benchmark, the estimations are within 7 % of ARM published data. To better illustrate the trade-offs, we also simulated more complicated workloads from PARSEC. It’s generally believed that the A7 is between 3 and 4x more energy efficient, while the A15 can speed-up execution up to 3x. Our framework correctly simulates these trends.

Figure 1: Simulation accuracy vs speed (with normalized computing resources).

Figure 2: Simulated energy and performance trade-offs of big.LITTLE CPUs.

Related Publications:
Physical attacks are a family of very effective attacks against embedded devices. They encompass three kinds of ways to recover information about the chip under attack: physical inspection, side channel analysis (listening to the activity of the chip in order to infer some hidden information) and fault attacks (putting the chip in unexpected or failure situations to recover information about its internal workings).

Physical attacks are generally split into two steps, thereafter called 'first step' attacks and 'second step' attacks. 'First step' attacks consist in a phase of 'reverse engineering': identifying the security functions implemented, when and where they are executed. Then, the 'second step' attack focuses on the potential weakness previously identified. These 'second step' attacks are very powerful but always rely on a precise spatial and temporal control of the target. In particular, the success of a fault (resp. side channel) attack depends on the time precision required to inject the fault (resp. measure the physical characteristics) at the right time and on the right area of the chip during program execution. Many protections (also called 'ad hoc') have been proposed to counter 'second step' attack schemes. But surprisingly, few protections have been proposed against 'first step' attacks, even if these ones constitute a crucial stepping stone for the success of an attack.

In this work, we study the effectiveness of runtime code polymorphism, which consists in regularly changing the behavior of a secured component at runtime while maintaining unchanged its functional properties, in order to make attackers' lives more difficult. Our hypothesis is that polymorphism can tackle security issues at several levels: it increases the difficulty of reverse engineering ('first step' attacks), it provides dynamically changing temporal and spatial properties that also increase the difficulty of 'second step' physical attacks. Furthermore, we assume that the use of polymorphic code is a realistic objective even for resource-constrained embedded devices, thanks to the technology developed in this work, which brings runtime code generation within reach of constrained embedded systems such as microcontrollers.

Figure 1 illustrates what an attacker is able to observe with low cost measurement hardware on an unprotected chip: an experienced attacker identifies the SubBytes function of AES and, with a DPA attack, can extract the cipher key with only a few dozens of measurements. Figure 2 illustrates the effect of runtime code polymorphism applied to the same AES cipher: the attack is no longer successful, even after ten thousands measurements. In this experiment, runtime code generation is achieved on a STM32 board fitted with an ARM Cortex-M3 core and only 8kBytes of RAM.

**ABSTRACT:** This work studies the interest of code polymorphism, achieved at runtime, as an effective protection against physical attacks. Code polymorphism is an effective means to increase the difficulty of reverse engineering, and to decrease the effectiveness of physical attacks (side channel attacks and fault attacks), which generally rely on particular physical or temporal properties of the target under attack. The technology developed in this work makes runtime code polymorphism within reach of constrained embedded systems such as microcontrollers.

**Related publications:**
Dynamic Compilation for embedded systems and HPC

V. Lomüller, H-P Charles

ABSTRACT: Dynamic compilation is a very powerful technology able to adapt binary code to new context. Unfortunately the general usage in JIT compilation does not take data sets into account and is not suitable for embedded systems (too big, too slow). We have developed tools and experiments that showed that dynamic compilation can be used in embedded systems and can be faster than traditional JIT in term of code generated and code generators. Our tool deGoal allows to generate fast code optimizers, and it’s possible to go even farther with a general tool, based on LLVM, called Kahuna. Both tools are 3 order of magnitude faster than classical JIT.

We have shown in [2] and [3] that it is possible to adapt running code to new applications context and improve performances. But performances are not the only interesting metric, especially in embedded systems. Memory footprint and code generation speed are also very important on systems which have very limited resources for power efficiency reasons.

The classical way to generate a binary code uses a compiler which translates the source code to the binary code. This scenario becomes less and less used, because: (1) developers prefers to use dynamic languages with powerful properties (object programming, scripting), (2) it becomes difficult to know the exact architecture characteristics (ARM has many architecture version, Intel has many SSE version, etc) and (3) micro architectures performances become more and more sensitive to data characteristics (data alignments for caches, data size for loop counts, data size for vectorization, etc).

The resulting data values are the output of the program. Depending on the characteristics of the data set (values, data size, etc) the performance results can vary a lot, especially on processors with caches or multimedia instructions.

We have developed the notion of "complettes" which are small code generators embedded into the application. Using a complette change the scenario: depending on data sets characteristics, the complette will generate the portion of the code that is critical for performances. Thanks to the knowledge of the data characteristics, the generated code is faster than the classical code and the generation can be very fast (3 order of magnitude faster and smaller than LLVM JIT, see below).

To generate complettes we have developed a tool called deGoal which is described in [3]. This tool allows very fast code generation, permits to use the full processor capabilities (parallel instructions, multimedia, etc) but needs a development effort. Victor Lomüller has developed another tool "Kahuna" based on LLVM. As shown in figure 2, deGoal and Kahuna are 3 orders of magnitude faster than LLVM JIT.

The general idea is illustrated figure 1. The usual way to generate code is on the left part. A compiler is in charge to transform a source code to a binary code. The binary program executed on a processor will process the data sets.
A compilation flow for parametric dataflow: Programming model, scheduling, and application to heterogeneous MPSoC

M. Dardaillon, K. Marquet and T. Risset (INSA-Lyon, CITI-INRIA), J. Martin, H.-P. Charles

ABSTRACT: Compared to synchronous dataflow (SDF), parametric dataflow models aim at providing sufficient dynamicity to model modern signal processing applications, while maintaining the high level of analyzability needed for efficient real life implementations. In this work we have designed a compilation flow for parametric dataflows. A specific attention is given to scheduling, to fill the crucial gap between dataflow models and real hardware on which actor firing is not atomic. Experimentations have been done by compiling 3GPP LTE-Advanced demodulation on a heterogeneous MPSoC with distributed scheduling features.

Our compilation flow, shown in Fig. 1, targets synchronous parametric dataflows (SPDF). Built on the LLVM compiler infrastructure, it offers an actor-based C++ programming model to describe parametric graphs, a compilation front-end providing graph analysis features, and a retargetable back-end to map the application on real hardware.

As shown on Fig. 2(a), input language benefits from C++ expressivity to ease description of complex applications. An intermediate on-the-fly execution of compiled code allows construction of the extended graph Fig. 2(b), which is then used for the back-end steps of compilation flow.

Unlike classical approaches, mapping of actors on a platform allows grouping several actors on the same processing core or IP, either because granularity of actors do not match IPs — in Fig. 3(b), 2 actors are simultaneously executed by OFDM1 IP — or because an IP may be used to execute several actors sequentially — e.g. SME3 for Fig. 3(c). These requirements are then taken into account by the scheduler.

To tackle the gap between SPDF model and non-atomic computation on real platform, we have introduced the notion of “micro-schedule”, which expresses the sequential order of input and output operations of each actor. These micro-schedules are then combined and merged depending on the mapping constraints, and micro-schedules of each processing cores are computed. Those include all necessary information to 1) formally check FIFO size, more precisely than with SDF, to guarantee absence of buffer-related deadlock in application execution, and 2) generate execution code for chosen hardware platform while taking benefit of any distributed scheduling feature it may offer.

Experiments have been run with a complex 4G-telecom application (3GPP LTE-Advanced) on the heterogeneous Magali MPSoC developed in CEA-Leti (Fig. 3). They have proven the relevancy of the approach, which may outperform manual optimizations that were done formerly.

Figure 1: Overview of compilation flow

As shown on Fig. 2(a), input language benefits from C++ expressivity to ease description of complex applications. An intermediate on-the-fly execution of compiled code allows construction of the extended graph Fig. 2(b), which is then used for the back-end steps of compilation flow.

Figure 2: Example of SPDF application

(a) Program excepti (reduced graph)  (b) Corresponding extended graph (for 4 antennas): SPDF model

Figure 3: Mapping of test case applications on CEA-Leti Magali chip.

Experiments have been run with a complex 4G-telecom application (3GPP LTE-Advanced) on the heterogeneous Magali MPSoC developed in CEA-Leti (Fig. 3). They have proven the relevancy of the approach, which may outperform manual optimizations that were done formerly.

Related Publications :
Designers of Embedded System (ES) face a number of difficult challenges. One is to have an efficient organization of teamwork, where several profiles can cooperate efficiently together. Another is the strong requirement on languages. Ideally, we would rely on a high-level language that: (1) Can model hardware as well as software; (2) Is not tied to any hardware architecture like Field-Programmable Gate Array (FPGA), Digital Signal Processor (DSP) or General Purpose Processor (GPP); (3) Is parallel-friendly; (4) Provides a clear path to generation of efficient code. Lastly, the wide variety of modern platform patterns like Globally Asynchronous Locally Synchronous (GALS), Network-on-Chip (NoC) based many-core platforms with distributed memory are more difficult to program than classical shared memory ones, and are increasingly heterogeneous. Parallelism has to be exposed and their distributed architecture requires strong partitioning of the code and calls for message-passing style of programming.

At the same time, modern embedded applications cannot be divided between data/computational parts and control parts anymore. Rather, they are made of a number of layers that include at the same time parallel computations on large data sets, as well as data-dependent control.

Based on an existing modeling framework, Highly Heterogeneous, Object-Oriented, Efficient Engineering (HOE2) [1], we propose a new parallel action language [2]. Each action is separated in two sequential sets of parallel events: first all the updates of the object associations, then all sending of messages to other objects.

Figure 1: HOE2 Action Language

Within this structure, we extended associations updates and send and receive primitives with indexed messages based on association’s multiplicity, to model data-parallel operations. Figure 2 describes indexed sending and receives primitives over high-level structures (Pixel). Index sets, colored in red, allow the programmer to properly set and recover index values. Although not illustrated here, the indexes can also be used in the part of the action clause that updates associations. This provide for array traversal and data permutation.

Figure 2: Indexed Messages

The proposed model presents a clean and simple expression for parallelism through indexed messages that is well suited for polyhedral analysis. Based on such analysis, messages and reactions can be matched at compilation time while preserving the parallel semantics.

We also present a compilation chain that builds on the idea of indexed messages a set of analyses and optimizations for efficient code generation. Most existing MDE frameworks rely on foreign action languages such as C/C++ to model actual computations, and the semantics of the original computational model (HSM) is lost in the process and hardly taken into account for optimization purposes by low-level languages.

By refining the model along the compilation flow, we look for final model patterns that are known to have efficient implementations in the target language.

Related Publications:
In hard real-time systems, the worst-case end-to-end delay of all the packets generated by a flow must be lower than a predetermined deadline. The goal of new methods for real-time packet schedulability analysis is to reduce the pessimism in the values that are obtained. In this work, we focus on wormhole networks where a packet is divided in flow control digits (flits) of fixed size, which are transmitted one by one by routers. The header flit, i.e. the first flit, contains the routing information that defines the path the next flits will follow in a pipeline way.

However, current methods to compute worst-case end-to-end delays in wormhole networks (i.e. RC method) do not model this pipeline way of transmitting flits. Instead, the delays of all the flows that block the analyzed flows are simply added. In this work, we show that integrating this pipeline behavior of wormhole networks in the computation of delays can greatly reduce the pessimism of the obtained values. We illustrate that on two examples and compare with the delay obtained by applying RC method.

Let us note $E_r$ the number of empty routers separating the analyzed flow and indirect flow. Let us further note $N_f$ the number of flits a packet is made of. Then, the condition to check whether an indirect flow can impact the transmission delay of the analyzed flow is the following: $E_r > N_f - 1$.

Figure 1: A simple NoC architecture

Fig.1 illustrates a simple NoC architecture and the flows we consider in our two examples. The analyzed flow is $f_1$. In our first example we only have the flow $f_2$ in direct contention with $f_1$, while in our second example we add the flow $f_3$ in indirect contention with $f_1$. In the RC method, the analyzed flow is considered to be blocked until all flows in contention reach their destinations. The overall delay of $f_1$ is equal to 47ns in a numerical application. We now show that this can be pessimistic. Fig.2 indeed shows the timeline of the transmission of the flits of our first example. When the third flit of $f_2$ crosses $r_2$ at time $t_1$, the header flit of $f_1$ is transmitted by $r_1$. In this case $f_2$ will no longer impact the delay of $f_1$. $f_1$ impacts a delay corresponding to $t_2 = 32$ns.

We now consider our second example. Fig.3 shows the timeline of the transmission, where it can be noticed that $f_3$ does not block $f_1$. Obviously, the delay of flow $f_1$ in this second example is identical to its delay in the first example. While using RC method, the delay of the indirect flow $f_3$ is further added, leading to a delay equals to 63ns for $f_1$. These delays are 1.5 and 2 times higher than the delays considering the pipeline behavior.

As shown in the examples given in this paper, only a part of the blocking flows should be considered in order to compute tighter end-to-end transmission delays, due to the pipeline behavior of the network. We also formulate a first property allowing to check if an indirect flow can impact the transmission delay of the analyzed flow.

Related Publications

In hard real time computer systems, correctness of a result depends on both the time and the value domains. With the increasing complexity of these systems, ensuring their correctness using a posteriori verification becomes, at best, a major factor in the development cost and, at worst, simply impossible. An error in the specifications is not detectable. We must, therefore, define a structured and simplified design process which allows the development of correct by-construction system.

Two fundamentally different paradigms for the design of real-time systems are identified; Event-Triggered (ET) and Time-Triggered (TT) paradigms. Activities in TT paradigm are initiated periodically at predetermined points in time. These statically defined activation instants enforce regularity and make TT systems more predictable than ET systems. This approach is well-suited for hard real-time systems. A system model of this paradigm is essential to speed-up understanding and smooth design task.

The three key-concepts of TT paradigm were identified:
(1) The global notion of time that must be established by a periodic clock synchronization in order to enable a TT communication and computation,
(2) The temporal structure of each task, consisting of predefined start and worst-case termination instants attributed statically to each task and
(3) TT interfaces which is a memory element shared between two interfacing subsystems.

In order to target a wider spectrum of TT implementations, we consider approaches for building component-based systems that provide a physical model from a high-level model of the system and TT specifications. In addition, if these approaches provide correctness-by-construction, they can help to avoid the monolithic a posteriori validation.

From a high-level BIP system model, a physical model containing all TT concepts (such as TT interfaces, the global notion of time and the temporal structure of each task) is generated using a set of source-to-source transformations. The output model of this transformation should have the structure depicted in Figure 2.

This physical model (called also BIP-TT model) is then translated to the programming language specific to the particular TT platform. The program in this language is then compiled using the associated compilation chain. Thus, BIP-TT model is not dedicated to an exemplary architecture.

First we have addressed the issue of source-to-source transformations that explicit TT communications in the physical model, in BIP framework.

An example of a BIP model is given in Figure 1. We consider the Real-Time BIP version where atomic components are represented by timed automata.

![Figure 1: Example of a BIP Model](image)

![Figure 2: Output model structure of the transformation](image)

**ABSTRACT:** To design a Time-Triggered (TT) system, we consider approaches for building component-based systems providing a physical model from a system high-level model and TT specifications. A source-to-source model transformation is applied to the initial model, taking into account key-concepts of TT paradigm. The obtained physical model is thus suitable for direct transformation into languages of specific TT platforms. In addition, if these approaches provide correctness-by-construction, they can help to avoid the monolithic a posteriori validation.

**Research topics:** component-based design, real-time systems, correct by construction
A Model of for Real-Time Applications on Embedded Many-cores

S. Louise, P. Dubrulle, T. Goubier

ABSTRACT: Manycore processors are the new trend of microprocessor industry to increase performance without impacting power consumption. Stream programming and Data-Flow based concepts have been successful in harnessing manycores incredible computing power for signal and image processing applications. Nonetheless, they lack the dependability and the versatility required for the emerging High Performance embedded applications like Autonomous vehicles and ADAS or Augmented Reality. We successfully enhanced the well-known CSDF model of computation with simple data processing actors that retain the good properties of CSDF.

Data-flow Models of Computations (MoC) are both simple and quite efficient for managing parallelism. They can be seen as a formalization of the old “processes and pipes” Unix-way of programming, as pipes are the only communication means between processes.

With the emergence of multi-core and many-core systems for which power consumption is mastered by utilizing more parallelism, there was a new interest in these data-flow models, especially in the embedded domain. Programming languages like StreamIt, and more recently the C (“sigma-C”) language [1], created by the CEA LIST, were overwhelmingly successful at harnessing the processing power of manycores in the case of signal and image processing applications. This success was enough to recreate also an interest of the HPC community to face the issues of the so-called “Exascale Challenge”, because large-scale memory coherence is foreseen as difficult to achieve. But beyond these good results, the progresses are reluctantly done for programming embedded systems. The reason is that, on one hand, data-flow Models of Computations (MoC) that have well behaved and have good properties like CSDF (for Cyclo-Static Data-Flow) lack the dynamicity and the versatility required for emerging embedded applications with time constraints, and on the over hand dynamic data-flows does not offer the required dependability, execution determinism and absence of deadlocks. To grant the power of developing massively parallel application with real-time constraints, a data-flow MoC would require both the expressiveness of dynamic data-flow and the determinism of CSDF.

This is the contribution we have done in this paper. To introduce the requirements, we based our analyses on several case studies extracted from characteristic parts of emerging applications. Such applications are not yet well handled by current programming models but they are believed to be important parts of the future of embedded systems. They include ADAS and autonomous vehicles, Augmented Reality and Cognitive Radios as preeminent figures of these emerging applications.

For real-time systems, liveness and timeliness are two important properties. In CSDF graphs, it is mathematically decisive if the said graph can be executed in bounded memory and without deadlocks. Therefore CSDF applications can be made globally deterministic and live which are 2 important properties of any enhanced MoC. But to ensure timeliness and liveness even when communication channels are fragile, it is required to go further. It is a known fact that it is possible to extend the semantic of data-flow graph by using special data-distribution actors like Splitters (round-robin distribution of data for SPMD parallelism), Duplicaters and Joiners (which collect data from several sources) [2].

For real-time systems, liveness and timeliness are two important properties. In CSDF graphs, it is mathematically decisive if the said graph can be executed in bounded memory and without deadlocks. Therefore CSDF applications can be made globally deterministic and live which are important properties of any enhanced MoC. But to ensure timeliness and liveness even when communication channels are fragile, it is required to go further. It is a known fact that it is possible to extend the semantic of data-flow graph by using special data-distribution actors like Splitters (round-robin distribution of data for SPMD parallelism), Duplicaters and Joiners (which collect data from several sources) [2].

To that end we introduced 2 new distribution actors: Duplicate-selectors and Transaction (with or without deadline) which are outside the CSDF formalism but can be combined in CSDF equivalent constructs (see e.g. Fig. 1). These data-distribution actors, once combined in valid constructs were shown to allows for managing fragile communication channels, open systems (IoT/CN) and account for redundancy (important for dependable systems), speculative execution (important to improve parallelism), and allow for easy deadline management including the difficult case of incremental evaluations (see Fig. 1). It can also be applied to manage power saving modes in such applications. The scope of all the possible applications remains to be fully explored [3].

Related Publications:
Emerging Technologies & Paradigms

- Neuro-Inspired IPs
- Resistive Memories
- Monolithic 3D
- Carbon Nanotube & Graphene
A Multiprocessor System-on-Chip (MPSoC) is built of multiple Processing Elements (PE) that can work in parallel, Fig. 1. Each PE or set of PEs form a Voltage/Frequency Island (VFI), i.e. they work within the same power domain. The supply voltage $V_{dd}$ and frequency $f$ are set by dedicated switching circuits allowing Dynamic Voltage and Frequency Scaling (DVFS). By decreasing the speed of the PEs with lower performance requirements the energy consumption is reduced. A control unit decides on $(V_{dd}, f)$ based on workload, latency, temperature, ... A dynamic approach such as game theory has been used to control PEs frequencies at runtime [1]. Recently Encoded Neural Networks (ENNs) have been used as the controller showing important gains in terms of time response and energy consumption [2].

In this work, the ENN is applied to a 4G telecom application on FAUST MPSoC. The test-case consists of five tasks mapped on five PEs. Each PE can modify its $(V_{dd}, f)$ couple. The objective of the controller is to minimize the total energy consumption of the system and preserve the global latency constraint $L$ to finish all the tasks. The control unit in Fig. 1 is replaced with a part of the ENN. The ENN is used to store $(V_{dd}, f)$ settings for each PE, precomputed for a number of possible $L$ values. Then, the $L$ value is fed to the ENN by a higher system layer and the $(V_{dd}, f)$ for each PE are output by the ENN.

First, the energy savings brought in by the ENN are compared to a method consisting in setting a global frequency to all the PEs to complete all the tasks at the latency $L$. Fig. 2 depicts the energy gain with respect to the global frequency as function of the possible number of latency $L$ values requested by the higher system layer [3]. Different curves correspond to a given number of $L$ values stored in the network. Fig. 2 shows how the gap between the possible and stored numbers of $L$ values impacts the energy gain.

Second, the comparison is done with the same scenario as in [1] which considers a latency $L=850\mu s$ and an energy budget of $280\mu J$. Since the game theory-based controller needs $105\mu s$ to converge [1], higher $(V_{dd}, f)$ have to be chosen to finish all the tasks before the latency constraint $L$. On the other hand, the convergence time of the ENN is negligible (27ns), and consequently lower $(V_{dd}, f)$ settings can be applied. In this scenario the game theory-controller saves 38% of energy budget while the ENN 60%. Furthermore, faster convergence also implies more efficient design. Since the amount of data waiting for the computation is smaller, memory buffers in the system can be reduced.

The proposed approach is the first one, to our knowledge, allowing a DVFS circuit to continuously adapt the power distribution of an MPSoC.

Related publications:
Recently Gripon and Berrou proposed a new model of neural networks that relies on a concept of neural clique and a cluster structure. This model is able to store a large number of messages and recall them, even when a significant part of the input is erased. The simulations of the network proved a huge gain in performance compared to state-of-the-art.

The network as presented by its authors is analyzed only for uniform independent identically distributed values among all the messages. This means that the number of connections going out from each node is uniformly distributed across the whole network. On the other hand, it is expected that real world applications may contain highly correlated data. Therefore, adapting the network to non-uniform messages is mandatory.

In order to store messages, we use a network that consists of binary neurons (nodes) and binary connections (Fig. 1). The network is divided in disjoint groups called clusters. In Fig. 1, nodes belonging to specific clusters are represented with different shapes. The connections are allowed only between nodes belonging to different clusters. Contrary to classical neural networks the connections do not have weights. To store a message (elementary part of information) in the network, one node in each cluster is chosen based on the content of the message. When the messages stored in the network have a non-uniformly distributed content, some of the nodes are chosen more often than others, which degrades the retrieval process and introduces additional errors.

Fig. 2 shows the error rate in function of the number of stored messages [1]. One can observe a degradation of the storage capacity in case of non-uniform Gaussian distribution. Among the proposed network adaptations, Huffman coding gives the best results. The core idea relies on the fact that Huffman coding produces variable-length codewords - the values that occur most frequently are coded on a small number of bits, whereas less frequent values are coded on a larger number of bits. Therefore, the sizes of the frequent values that break the uniformity are minimized. The free space is filled with random uniformly generated bits. Consequently, the most often appearing values are associated with the largest number of randomly chosen nodes and the influence of local high density areas is minimized. In the example in Fig. 2, the curve for Huffman coding reaches 0.1 error rate for 117000 stored messages.

In order to apply these recently introduced memories in practical applications, the proposed adaptations are necessary, as real-world data is not necessarily uniformly distributed. The obtained results are promising in terms of applications.

**Related Publications:**
The Highly Appropriate Marriage of 3D Stacking and Neuromorphic Accelerators

Research topics: Neural network, 3D stacking, accelerator

A. Valentin, B. Belhadj (INRIA), P. Vivet, M. Duranton, L. He (INRIA), O. Temam (INRIA)

ABSTRACT: 3D stacking is a promising technology (low latency/power/area, high bandwidth); its main shortcoming is increased power density. Simultaneously, motivated by energy constraints, architectures are evolving towards greater customization, with tasks delegated to accelerators. Due to the widespread use of machine-learning algorithms and the re-emergence of neural networks (NNs) as the preferred such algorithms, NN accelerators are receiving increased attention. They turn out to be well matched to 3D stacking: inherently 3D structures with a low power density and high across-layer bandwidth.

Power consumption and dissipation limitations drive a strong trend towards heterogeneous multi-cores, composed of a mix of cores and accelerators. Accelerators can be highly energy-efficient because they are specialized circuits, but at the same time they have, almost by definition, a limited application scope. In that context, neuromorphic accelerators have several major assets. They are suitable for major categories of applications: machine-learning algorithms are at the heart of many emerging applications, especially on the client (e.g., voice recognition with Siri, object recognition with Google Goggles, etc). The fact that one kind of algorithm works well across many applications means there is a unique opportunity to design a circuit with ASIC-like efficiency but with a broad application scope. As a result, in the past few years, researchers have started to investigate neuromorphic accelerators. Furthermore, it is a matter of fact that Neural networks (NN) marry well with 3D technology. First, modern NNs differ from the models popular in the 1990s by their large number of layers, leading to the term Deep Neural Networks (DNNs). Second, one NN layer is often a 2D structure, especially for image recognition applications, so that a NN is often an inherently 3D structure. Mapping a 3D structure onto 2D circuits results in either multiple long wires between layers or congestion points. Third, computations in NNs are parallel and almost evenly distributed among neurons; this prevents the occurrence of hotspots in 3D implementation, contrary to high clock frequency circuits for which power density in 3D becomes an issue.

The NN accelerator is designed here for image classification in an industrial application: the aim is to be able to pinpoint defective parts moving on a conveyor belt. The accelerator will be connected to a silicon retina, which generates spikes as function of the objects moving in front of it. The image sensor contains 48 macro-blocks, of 16x16 8-bit pixels each. The neural network is composed of two layers (see Fig. 1). The first one is used for extracting simple shapes (horizontal, vertical, diagonal segments). Its neurons therefore do not need to see the entire image but a subpart of it. ‘First layer’ connectivity is inspired by biology, with the concept of ‘local receptive fields’: each neuron sees four 16x16 pixels macro-blocks, leading to 1024 associated synaptic weights (coded on 8 bits). The 48 neurons add incoming spikes in their soma. It is implemented as a 3-stage pipeline, composed of: (1) a Four input adder, to add simultaneously 4 spikes from 4 macro-blocks; (2) a register, to store the soma value; (3) a comparator, to generate a spike if the soma value goes above threshold. The ‘Second layer’ is a classifier. It is composed of 64 Neurons: the one with the strongest response is deemed the winner, i.e. it gives the shape/object that was more strongly recognized. This layer needs a full connectivity on the entire image. Therefore, each neuron is connected to all 48 neurons of ‘First layer’. An adder tree is used to add the 48 corresponding 8-bit synaptic weights, in an 8-stage pipeline. The two layers are physically implemented in different tiers, in 130nm technology. They are connected thanks to wafer bonding: power delivery and signal propagation is ensured thanks to micro-bumps of 5x5um².

Figure 1. Schematic of a 2-layer neural network for vision processing

This 3D implementation is compared to a 2D one, in Table 1. The power and frequency gains are mostly owing to shorter interconnect wires.

Table 1. Comparison of the 2D and 3D implementations of the NN accelerator

<table>
<thead>
<tr>
<th></th>
<th>2D circuit</th>
<th>3D circuit</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical path (ns)</td>
<td>9</td>
<td>6.63</td>
<td>-26%</td>
</tr>
<tr>
<td>Power (mW)</td>
<td>428.24</td>
<td>353.9</td>
<td>-17%</td>
</tr>
<tr>
<td>Area (mm²)</td>
<td>7.974</td>
<td>3.634</td>
<td>-54%</td>
</tr>
<tr>
<td>Wires (m)</td>
<td>19.9</td>
<td>15.6</td>
<td>-21%</td>
</tr>
</tbody>
</table>

Related Publications:
In this work, we compare different CBRAM addressing structures for use in synaptic arrays. Crossbar (1R), anode connected matrix (1T-1R), where the gate of each selection transistor on a same anode are connected together and cathode connected matrix (1T-1R), where the gate of each selection transistor are connected together. The crossbar structure is known to be an energy consuming structure due to sneak paths. It has other drawbacks such as read current variation and neighboring cells induced switching.

The memristive device simulated in this work is from the CBRAM family. This technology is non-volatile and bipolar: a positive voltage across the device terminals will turn the device ON in a low resistance state (LRS) and SET the device. A negative voltage will turn the device OFF in a high resistance state (HRS) and RESET the device. To simulate the CBRAM, we used a physics-based model using VHDL-AMS based on work by [Yu, 2011] and simulated with "Spectre 7.1.1." We improved the model by bringing continuity in the CBRAM resistance calculation, which allows taking into account the current compliance intrinsically in the model.

To calculate the voltage drop for larger crossbars, we simplified it by eliminating path where current are negligible, resulting in the schematic in Figure 1-a). By using recurrent Y - Δ transform, it is possible to iteratively simplify the network to a final Y topology with equivalent resistances as shown in Figure 1-b). From this schematic, the actual voltage across the farthest CBRAM VCBRAM is retrieved using Millman’s theorem.

To avoid programming neighboring cells, an intermediate voltage source Vi is introduced that can be applied to the inactive cathodes/ anodes. We reproduced the worst case scenario by setting all CBRAM in LRS at 4.5 K. In this setup, 1.32 mA get dispersed into the crossbar during a SET where only 331 µA are actually required to switch the selected device.

Figure 2 shows the actual SET voltage on the farthest CBRAM vs matrix size for the anode connected structure. For an "ON" resistance of 4.5 K ohm, a matrix (1T-1R), of size up to 20,000x20,000 can be achieved as compared to 150x150 in the case of a simple (1R) crossbar.

In this work, we modelled and simulated specific issues related to the CBRAM technology such as incomplete SET switching due to voltage drops across anode and cathode lines as well as effect of neighboring cells induced switching. We found out that the anode connected structure is ideal during a read mode but it requires sequential programming while the cathode connected structure is better suited for use in an on-line programming neural network.

Related Publications:
Circuits and systems for the Internet of Things are designed to operate for years without human intervention. These devices spend more than 90% of their lifetime in sleep state. Suppressing the power consumption during these phases would considerably increase the battery lifetime of wireless systems.

The easy co-integration of emerging Non-Volatile memories in advanced CMOS process opens the way of distributed non-volatility and increases the power efficiency of context saving. Thus, our work focuses on embedding the non-volatility in one of the key elements of digital circuits: the Flip-Flop.

The oxide-based RRAM technology (OxRAM) appears as a key enabling technology for low cost applications due to its technological advantages (back-end-of-line integration without process modifications, fast switching and high scalability).

Several Non-Volatile Flip-Flop architectures are reported [1]. These solutions include the non volatile device directly inside the Flip-Flop causing huge modifications of the architecture. We proposed an architecture of Non-Volatile Flip-Flop which requires no modifications of the Flip-Flop itself and uses OxRAM to save the context.

The proposed architecture consists in tying externally a Non-Volatile Latch to a regular Flip-Flop (Fig.1.).

Adding non-volatility allows storing the flip-flop content on the NV part, completely cutting off the power supply during the long sleep periods and recovering their content during wake-up. Due to supply power off, regardless of the sleep time, the power consumption is constant (Fig.2.). Owing to external addition of the NV feature, during the active mode the flip-flop operates without any performance degradation.

Adding non-volatility allows storing the flip-flop content on the NV part, completely cutting off the power supply during the long sleep periods and recovering their content during wake-up. Due to supply power off, regardless of the sleep time, the power consumption is constant (Fig.2.). Owing to external addition of the NV feature, during the active mode the flip-flop operates without any performance degradation.

Programming circuit of the OxRAM memory is designed to operate at 1.5 Volts, the best trade off between energy consumption and FDSOI 28nm thin gate oxide transistors reliability requirements. Active and restore modes operate down to 0.4V of supply voltage.

The proposed non volatile Flip-Flop design provides a simple standard-cell based non-volatile latch solution fully compatible with thin oxide digital design. We demonstrated a low HRS/LRS required ratio for restore operation which ensures good design margins against OxRAM variability. Near-threshold voltage restore operation enables ultra-wide voltage range NVFF design. Low penalty delay makes the NVFF competitive with balloon latch.

Figure 1: Non Volatile Flip-Flop With detailed operating modes

Figure 2: Sleep mode energy for Non-Volatile Flip-Flop, compared to several retention Flip-Flops at different supply voltages.

The proposed non volatile Flip-Flop design provides a simple standard-cell based non-volatile latch solution fully compatible with thin oxide digital design. We demonstrated a low HRS/LRS required ratio for restore operation which ensures good design margins against OxRAM variability. Near-threshold voltage restore operation enables ultra-wide voltage range NVFF design. Low penalty delay makes the NVFF competitive with balloon latch.

Related Publications:

ABSTRACT: The increasing amount of circuits and systems operating on battery and energy harvested sources highlight the need for ultra-low standby power. Including Non-Volatility inside the Flip-Flops allows nullifying the power consumption during sleep mode phases while keeping the system state. The proposed solution consists in a simple integration of OxRAM memories inside a Flip-Flop to provide context saving and restore, before and after a sleep phase, In 28nm CMOS FDSOI, adding non-volatility cut-off the Flip-Flop leakage at the cost of 63pJ of energy.
3DCoB: A new design approach for Monolithic 3D Integrated Circuits

Research topics: Monolithic 3D Design, Cell-on-Cell, NoverP, Cell-on-Buffer

Hossam Sarhan, Sebastien Thuries, Olivier Billoint and Fabien Clermidy

ABSTRACT: 3D Monolithic Integration (3DMI) technology provides very high dense vertical interconnects with low parasitics. Previous 3DMI design approaches provide either cell-on-cell or transistor-on-transistor integration. In this paper we present 3D Cell-on-Buffer (3DCoB) as a novel design approach for 3DMI. Our approach provides a fully compatible sign-off physical implementation flow with conventional 2D tools. We implement our approach on a set of benchmark circuits using 28nm-FDSOI technology. The sign-off performance results show 35% improvement compared to the same 2D design.

3D integration becomes an emerging technology to overcome the limitations of scaling of the advanced technology nodes. 3D integrated circuits can be fabricated by two main process techniques: parallel integration and sequential integration.

Parallel integration is achieved by fabricating two dies separately then vertically stacking using 3D integration technologies such as 3D Through-Silicon-Vias (TSVs) or 3D interposers [1].

On the other hand, sequential integration like monolithic 3D integration (M3D) is a methodology to fabricate the 3D IC where a second transistor layer is directly fabricated on top of the first one. The two transistors layers are connected through high-density inter-tier vias (about 100 times smaller compared to the TSVs [2]).

At design level consideration, parallel integration uses the conventional 2D standard cells however it suffers from the incompatibility of the 2D design tools. On the other hand, the transistor-level integration is achieved by splitting each standard cell between two tiers (NMOS transistors are placed in one tier and PMOS transistors are placed in another tier). Transistor-level integration provides compatible approach with the 2D conventional physical implementation tools but it requires re-designing of the standard cells layout.

A new design approach is introduced here for monolithic 3D integration providing a full compatibility with the conventional 2D sign-off physical implementation flow. In addition it improves the performance compared to a 2D design as shown in figure 2 (up to 35% better performances compared to 2D without extra power penalty).

The proposed idea is to split the non-minimum drive standard cell into a logic stage and a driving stage. The logic stage is implemented by the equivalent minimum-drive cell. The driving stage is implemented by a buffer with the same drive as the original cell. The min-drive cell and the driving buffer are vertically stacked which we call 3D Cell-on-Buffer (3DCoB). Using this approach, the minimum-drive logic cell provides the same logical function as the original cell, while the driving buffer guarantees the same driving capability of the cell as shown in figure 1.

Table 1 summarizes the comparison of the 3D Design characteristics.

<table>
<thead>
<tr>
<th></th>
<th>N/P</th>
<th>Cell-on-Cell</th>
<th>3DCoB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D Design flow compatability</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Using 2D standard cells</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Using inter-tier routing metal layers</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Usage of inter-tier vias</td>
<td>In every cell</td>
<td>Between cells (if req.)</td>
<td>Only in the 3D cells</td>
</tr>
</tbody>
</table>

Table 1: Comparison of M3D Design Characteristics [3]

Figure 1: M3D CoB Approach [3]

Figure 2: 2D vs 3DCoB Performances Analysis for AES 128 block at Area = 0.119 mm²[3]
Accurate models for high-speed analog and RF simulation of carbon nanotube and graphene FETs

Research topics: Carbon nanotube FETs, Graphene FETs, RF integrated circuits

José Luis González, Gerhard Martin Landauer (UPC Barcelona-Tech*)

ABSTRACT: Carbon based devices are proposed as good candidates for beyond CMOS technologies. In this work, we have developed new models for carbon nanotubes and graphene FETs suitable for high-speed analog and RF simulation that are required for the assessment of such devices for circuit design.

The evaluation of the suitability of carbon nanotube FETs (CNT-FET) and graphene FETs (G-FET) for high-speed and RF applications relies on device models that capture the most important effects of the devices with respect to such applications, i.e. variability, noise, and extrinsic parasitics. Moreover, the models should accurately predict important parameters such as small signal gains and noise figures. Previous models found in the literature do not fulfill such requirements.

In the framework of a three-year research project, we have developed new models for CNT-FETs and G-FETs aimed at high-speed analog simulations. The CNT-FET model (see Fig. 1 left) includes non-uniform tube diameter distribution inside the multi-finger FET and metallic removal process non-idealities [1]. It also features specific noise source models for CNT devices, such as suppressed shot-noise [2]. The G-FET model [3] includes a more accurate modeling of the quantum capacitance and improved accurate for biasing points close to the Dirac point of its band structure. Both models are implemented in Verilog A and are coded so that it can be used in the framework of commercial IC design and simulation tools. They feature a parametric model of the extrinsic device regions (see Fig. 1 right for the G-FET case), which is required for correctly obtaining the maximum operating frequency of such devices.

Figure 1 : Left: model of the CNT-FET transistor including nanotubes distribution non-idealities such as diameter dispersion and metallic tubes removal imperfections. Right: cross-section of the graphene-FET device showing the intrinsic and extrinsic sections.

The CNT-FET model has been used for designing and optimizing sub-THz low-noise amplifiers (LNAs) and Oscillators (VCOs), which compare very favorably with state of the art CMOS circuits (see Fig. 2). The model allows for investigating the impact of the non-uniformities in the tube diameter distribution inside the device on the dispersion of the RF figure of merit and accurately predicts important figures of merit such that the LNA noise figure or the VCO phase noise.

Figure 2 : Figures of merit for subTHz VCOs realized with various CMOS technologies and with CNT-FETs

The important improvements introduced in the G-FET model removes the uncertainty existing in previous models for low bias values around the Dirac point of the graphene single-layer that constitutes the device channel. For example, Fig. 3 shows that previous models introduced a large error when predicting the transconductance parameter (gm) for bias voltages around 0.

Figure 3 : Comparison of the new G-FET model with experimental data and previous models for the transconductance parameter.

The new models are fundamental tools for exploring the new possibilities that these devices offer to analog and RF circuit designers and allow benchmarking those two carbon-based technologies in comparison with other alternatives such as deeply scaled CMOS FDSOI or FinFETs.

*G.M. Landauer is now at NXP-Graz

Related Publications:
PhD degrees awarded in 2014

Wafa Ben Hassen
David Roclin
Vincent Legout
Mathieu Texier
Clément Jany
José Moron Guerra
Amine Didioui
Timothé Laforest
Fadoua Guezzi Messaoud
Romain Grezaud
Florent Dupont
Bertrand Pelloux Prayer
Bernard Sébastien
Ogun Turkyilmaz
Yassine Fkih
Yeter Akgul
Victor Lomuller
Nicolas Hili
Ozan Gunalp
Mickael Dardaillon
Study of embedded diagnosis strategies in complex wired networks

This study addresses embedded diagnosis of complex wired networks. Based on the reflectometry method, it aims at detecting and locating accurately electrical faults. Increasing demand for on-line diagnosis has imposed serious challenges on interference mitigation. It aims at making diagnosis while the target system is running. The interference becomes more critical in the case of complex networks where several reflectometers are injecting their test signals simultaneously. The objective is to develop new embedded diagnosis strategies in complex wired networks that would resolve interference problems and eliminate ambiguity related to the fault location.

The first contribution is the development of a new method called OMTDR (Orthogonal Multi-tone Time Domain Reflectometry). It uses orthogonal modulated digital signals for interference mitigation and thereby on-line diagnosis.

For better coverage of the network, the second contribution proposes to integrate communication between reflectometers. It uses sensors data fusion to facilitate decision making.

The third contribution addresses the problem of the diagnosis strategy, i.e. the optimization of diagnosis performance of a complex network under operational constraints. The use of Bayesian Networks allows us to study the impact of different factors and estimate the confidence level and thereby the reliability of the diagnosis results.

Integration of memory nano-devices in image sensors processing architecture

By using learning mechanisms extracted from recent discoveries in neuroscience, spiking neural networks have demonstrated their ability to efficiently analyze the large amount of data from our environment. The implementation of such circuits on conventional processors does not allow the efficient exploitation of their parallelism. The use of digital memory to implement the synaptic weight does not allow the parallel reading or the parallel programming of the synapses and it is limited by the bandwidth of the connection between the memory and the processing unit. Emergent memristive memory technologies could allow implementing this parallelism directly in the heart of the memory.

In this thesis, we consider the development of an embedded spiking neural network based on emerging memory devices.

First, we analyze a spiking network to optimize its different components: the neuron, the synapse and the STDP learning mechanism for digital implementation. Then, we consider implementing the synaptic memory with emergent memristive devices. Finally, we present the development of a neuromorphic chip co-integrating CMOS neurons with CBRAM synapses.
Energy-aware real-time scheduling of multiprocessor embedded systems

Real-time systems are often used in embedded context. Thus reducing their energy consumption is a growing concern to increase their autonomy. In this thesis, we target multiprocessor embedded systems which are becoming standard and are replacing uniprocessor systems. We aim to reduce the energy consumption of the processors which are the main components of the system. It includes both static and dynamic consumption and it is now dominated by static consumption as the semiconductor technology moves to deep sub-micron scale. Existing solutions mainly focused on dynamic consumption.

In this thesis, we propose new optimal multiprocessor real-time scheduling algorithms. They optimize the duration of the idle periods to activate the most appropriate low-power states. We target hard real-time systems with periodic tasks and also mixed-criticality systems where tasks with lower criticalities can tolerate deadline misses. Offline, we use an additional task to model the idle time and we use mixed integer linear programming to compute a schedule minimizing the energy consumption. Online, we extend an existing scheduling algorithm to increase the length of the existing idle periods. To the best of our knowledge, these are the first optimal multiprocessor scheduling algorithms minimizing the static energy consumption.

Evaluations have been performed using existing optimal multiprocessor real-time scheduling algorithms. Results show that the energy consumption while processors are idle is up to ten times reduced with our solutions compared to the existing multiprocessor real-time scheduling algorithms.

Dynamic Parallelism Management in Multi-Core Architectures for Mobile Systems

The amount of smartphone sales recently surpassed the desktop computer ones. This is mainly due to the smart integration of several functionalities in the same architecture. This is also due to the wide variety of supported applications like augmented reality, video conferencing and video games. The support of these applications is made by heterogeneous computing resources specialized to support each application type thus allowing to meet required performance and power consumption.

The objective of this work is to optimize the usage of all available computing resources at runtime. The selected use case is a graphic rendering pipeline application because it is a dynamic application, which is widely used in mobile devices. The application has been implemented and parallelized on a multicore architecture simulator. The profiling shows that the best balance of the parallelism depends on the rendered scene; a dynamic load balancing is therefore required for this application. The system has been implemented in a Timed-TLM simulator in order to estimate performance improvements allowed by the dynamic adaptation. An architecture to allow the acceleration of mobile applications has been defined and compared to other multicore architectures. The hardware complexity and the performance of the architecture have also been estimated. For an increased complexity lower that 1.5%, we demonstrate performance improvements up to 20% compared with static parallelisms. We also demonstrated the ability to support a variable amount of resources.
PhD degrees awarded in 2014

Clément Jany
University: Université Grenoble Alpes

Design and study of an innovative frequency synthesis in advanced CMOS technologies for millimeter-wave applications

The 60-GHz unlicensed band is a promising alternative to perform the high data rate required in the next generation of wireless communication systems. Complex modulations such as OFDM or 64-QAM allow reaching multi-gigabits per second throughput over up to several tens of meters in standard CMOS technologies. This performance rely on the use of high performance millimeter-wave frequency synthesizer in the RF front-end. In this work, an original architecture is proposed to generate this high performance millimeter-wave frequency synthesizer. It is based on a high order (several tens) multiplication of a low frequency reference (few GHz), that is capable of copying the low frequency reference spectral properties. This high order frequency multiplication is performed in two steps. Firstly, a multi-harmonic signal which power is located around the harmonic of interest is generated from the low frequency reference signal. Secondly, the harmonic of interest is filtered out from this multi-harmonic signal. Both steps rely on the specific use of oscillators. This work deals with the circuit design on advanced CMOS technologies (40 nm CMOS, 55 nm BiCMOS) for the proof of concept and on the theoretical study of this system. This novel technique is experimentally validated by measurements on the fabricated circuits and exhibit state-of-the-art performance. The analytical study of this high order frequency multiplication led to the discovery of a particular kind of synchronization in oscillators and to approximated solutions of the Van der Pol equation in two different practical cases. The perspectives of this work include the design of the low frequency reference and the integration of this frequency synthesizer in a complete RF front-end architecture.

José Moron Guerra
University: Université Lille I

Design of Sub-THz heterodyne receivers in 65 nm CMOS process

The main goal of this thesis is to explore design opportunities beyond the millimeter wave frequencies and to get as close as possible to the THz band using CMOS technologies. The main application is the heterodyne detection for THz imaging. The cut-off frequencies ft/fmax of the used process (65 nm CMOS) are 150/205 GHz, the chosen operation frequency of the developed systems is 280 GHz which means that the circuits developed during this thesis operate at least 80 GHz beyond their fmax cut-off frequency. Two 280 GHz sub-harmonic injection locked oscillators were developed; the injection frequency corresponds to one sixth of the output frequency. In order to generate oscillations beyond fmax, harmonic boost techniques are used such as the push-push and triple push techniques. The output powers of the oscillators are -19 and -14 dBm at 280 GHz. Both components were used as local oscillators for two heterodyne receivers operating around the same frequency. In order to down-convert the Sub-THz signal, a passive resistive mixer is used; this kind of circuit allows mixing beyond the active transistor limits. Also there is no LNA at the beginning of the Rx chain since the cut-off frequencies are very low and there will be no gain for amplification at 280 GHz. The conversion gain of both receivers is -6 dB however the NF's are 36 dB and 30 dB. The best receiver (30 dB) is co-integrated with an antenna (developed by Labsticc) using the same process allowing heterodyne detection THz imaging.
Energy-Aware Transceiver for Energy Harvesting Wireless Sensor Networks

Technological advances achieved in the fields of microsystems and wireless communications have enabled the development of small size and low cost sensor nodes able to establish a wireless sensor network (WSN). Many applications usually impose stringent constraints on the WSN lifetime which is expected to last several years. To reach this objective, it is necessary to reduce the overall energy consumption of the sensor node and to find an additional source of energy as well. To address the last point, energy harvesting from the environment seems to be an efficient approach. However, energy harvesting devices are usually unable to ensure a continuous operation of sensor nodes. Thus, it is necessary to adapt the WSN consumption and activity to the low and unpredictable energy scavenged.

The work presented in this thesis focuses on the issue of simulation and power consumption of autonomous sensor nodes. We have first developed, HarvWSNet, a co-simulation framework combining WSNet and Matlab that provides adequate tools to accurately simulate heterogeneous protocols and energy harvesting systems. We have demonstrated that HarvWSNet allows a rapid evaluation of energy-harvesting WSNs deployment scenarios. Thanks to the accurate energy models (battery, supercapacitor, etc.) implemented in this platform, we have studied and evaluated a large scale deployment of solar and wind energy-harvesting WSNs. Our second contribution focuses on the implementation of energy-aware reconfiguration strategies in the radio transceiver. These strategies are intended to reduce the excessive power consumption of the radio transceiver when the channel conditions are favorable. To this end, we have a new simulation framework called EnvAdapt (based also on WSNet). In EnvAdapt, we have implemented the required radio transceiver behavioral and power consumption models that allow the evaluation of the impact of radio transceiver reconfiguration on the communication performance and lifetime of WSNs.

Study of integrated devices for coherent light analysis and control: co-design of optoelectronic integrated circuits and optical systems

Among the optical medical imaging techniques used in medicine, the main limitation is the low resolution at a penetration depth greater than few mm. This limitation does not allow competing with the standard imaging techniques such as X rays or RMI based imaging. In that scope, the acousto-optical imaging features several advantages: it allows measuring an optical contrast useful to detect tumors, in conjunction with the spatial resolution of ultrasound. However, the state of the art detecting devices feature a lack of sensitivity, which prevent its transfer to medical practitioners.

This leads us to study the intrinsic features of the acousto-optical signal in order to propose two CMOS pixel architectures. The first one, fully analog, is compliant with the correlation time of biological tissue (1 ms typ.) and features an analog processing of the relevant signal. The second one is based on a digital pixel which contains an analog to digital converter, allowing simplifying the optical setup and increasing the robustness of the processing.

In addition, related to the recent progress in wavefront control, an opto-electronic device, coupled with the first pixel architecture, has been proposed. It allows performing an optical phase operation (e.g. phase conjugation) in parallel on a pixels array, within the correlation time of biological media. Thus, this monolithic device circumvents the speed limitations of state of the art setup by a physical stacking of a liquid crystals spatial light modulator over a CMOS image sensor.
Benefits of tri-dimensional integration for CMOS image sensors: Case study of high dynamic range imagers

This work leads to study the potential benefits of 3D integration for imagers. Extending the dynamic range of image sensors by self-adjustment of integration time is one way to guarantee the best use of the photodiode capacitance. We then choose this application to take advantage of dense vertical interconnections, to locally adapt the integration time by group of pixels, while keeping a classic pixel architecture and hence a high fill factor. Stacking the pixel section and the circuit section enables compact pixel, higher image quality and faster speeds by relaxing size constraints and count for pixel transistors. Moreover it allows the integration of flexible and versatile functions such as data compression. The proposed technique is based on adjusting the integration time by using a floating coding with a common exponent per group of pixels. The algorithm performs a first level of compression. A second level of compression is proposed, based on Discrete Cosine transform (DCT). It leads to reduce the data rate at the chip output. Higher compression ratios can be obtained, but at 95%, the result image is at a limit of being usable with a PSNR of 30dB. The proposed architecture to extend the imager dynamic range was been implemented and designed. A 2D HDR image sensor prototype was carried out in 180 nm CMOS Tower technology. The time integration is adapted at the cost of some pixels of each group of pixels sacrificed to implement necessary electronics for generating the HDR control signal. The test results are very promising proving the benefits of 3D integration notably in terms of power consumption and image quality. The use of 3D integration for a future generation of the proposed architecture will make possible the implementation of short feedback, due to vertical interconnections, to adjust the integration time without scarifying pixels and so improve image quality. It can also reduce the latency and the power consumption.

A Gate Driver for Diode-Less Wide Band Gap Devices-Based Synchronous Converters

Wide band gap devices already demonstrate static and dynamic performances better than silicon transistors. Compared to conventional silicon devices these new wide band gap transistors have some different characteristics that may affect power converter operations. The work presented in this PhD manuscript deals with a specific gate drive circuit for a robust, high power density and high efficiency wide band gap devices-based power converter. Two critical points have been especially studied. The first point is the higher sensitivity of wide band gap transistors to parasitic components. The second point is the lack of parasitic body diode between drain and source of HEMT GaN and JFET SiC. In order to drive these new power devices in the best way we propose innovative, robust and efficient solutions. Fully integrated gate drive circuits have been specifically developed for wide band gap devices. An adaptive output impedance gate driver provides an accurate control of wide band gap device switching waveforms directly on its gate side. Another gate drive circuit improves efficiency and reliability of diode-less wide band gap devices-based power converters thanks to an auto-adaptive and local dead-time management.
PhD degrees awarded in 2014

Florent Dupont
University: Université Grenoble Alpes

Impedance measurement integration into a multi-application implantable electrical stimulation system: proposition of an innovative stimulation strategy

This thesis involves the research, design, development and assessment of a novel electro-stimulation device capable of taking into account, and compensate for, the distortions of the therapeutic signal due to the presence of the electrode-electrolyte interface, thus ensuring that the desired signal waveform is in fact transmitted to the targeted tissues. This should result in more effective and energy-efficient stimulation and also enable fundamental research on the effects of signal waveforms on excitable tissue stimulation. The modeling of the electrode-tissue system enabled the demonstration that the electrical stimuli distortion, due to frequent filtering or stimulator saturation, can be avoided by using data extracted from a prior impedance spectroscopy. An equivalent electrical circuit is used to fit those data enabling the separation of electrode-electrolyte interface and tissue components. Based on the equivalent circuit, a transfer function is then defined and used in order to ensure the delivery of the user defined waveform to the stimulation targets. A proof of concept has been achieved with three steps: first on electronic components representing an in vivo impedance, then into a saline solution and finally into an actual animal for in vivo validation. Those tests led to functional experiments proving the interest of this novel method. The work is concluded by the specification and the simulation of an innovative multi-applications integrated architecture using impedance data in order to adapt stimuli waveforms.

Bertrand Pelloux Prayer
University: Université Grenoble Alpes

28-14nm UTBB FD-SOI design for low power and high performance digital applications

To meet the needs of mobile devices, combining high performances and low power consumption, the planar fully depleted silicon-on-insulator (FD-SOI) technology appears to be a suitable solution. Indeed, thanks to its thin-film of silicon and an undoped channel, MOS transistors have an excellent short-channel electrostatic control and a low variability of the threshold voltage given by an immunity to random dopant fluctuation. In addition, this compelling technology enables to adjust the threshold voltage of transistors by applying a wide ±3V back-bias voltage on Wells. Thus, this specific FD-SOI feature brings to IC designers an additional lever to modulate the performance and to optimize the energy efficiency of circuits.

The research work presented in this thesis has contributed to the development of FD-SOI technology platform for the 28 and 14nm nodes. Initially, a critical path extracted from an ARM Cortex-A9 processor was used to assess both the intrinsic gains provided by the FD-SOI technology and those produced by modulating the back-bias voltages. This technique enables to divide by up to 50 times the static current of circuits in standby mode, or by 2 the total energy consumption at same frequency. In the second phase, several design solutions are proposed in order to optimize the energy efficiency of circuits again. Among these, the design of a single-Well structure enables to solve the conventional multi-VT co-integration issue, occurring in FD-SOI. Moreover, this novel approach also offers an efficient solution for integrated circuits operating over a wide supply voltage range. Indeed, thanks to a single back-bias voltage, both n and p-MOS transistors could be easily balanced enabling an outstanding minimal supply voltage.

In modern ASIC designs, a large portion of the total power consumption is due to the clock network contribution, especially the leaves of the clock tree: the flip-flops. Therefore, the appropriate flip-flop architecture is a major choice to reach the speed and energy constraints of mobile and ULP applications.

After a thorough overview of the literature, the explicit pulse-triggered flip-flop topology is pointed out as a very interesting FF architecture for high-speed and low-power systems. However, it is today only used in high-performances circuits mainly because of its poor robustness at ultra-low voltage (ULV).

In this work, a large comparison of resettable and scannable latch architecture is performed in the energy-delay domain by modifying the sizing of the transistors, both at nominal and ultra-low voltage. Then, it is shown that the back biasing technique allowed thanks to the FDSOI technology provides better energy and delay performances than the sizing methodology.

As the pulse generator is the main cause of functional failure, we proposed a new architecture which provides both a good robustness, allowing correct operations at ULV. Moreover, this architecture permits a fine tuning of the delay and is so adaptable to a wide range of latches characteristics and applications.

Finally, a selected topology of explicit pulse-triggered flip-flop was implemented in a 16x32b register file. The speed, energy consumption and area performances are better than an implementation with master-slave flip-flops, mainly thanks to the sharing of the pulse generator over several latches.

Emerging 3D technologies for efficient implementation of FPGAs

The ever increasing complexity of digital systems leads the reconfigurable architectures such as Field Programmable Gate Arrays (FPGA) to become highly demanded because of their in-field (re)programmability and low nonrecurring engineering (NRE) costs. Reconfigurability is achieved with high number of point configuration memories which results in extreme application flexibility and, at the same time, significant overheads in area, performance, and power compared to Application Specific Integrated Circuits (ASIC) for the same functionality. In this thesis, we propose to design FPGAs with several 3D technologies for efficient FPGA circuits. First, we integrate resistive memory based blocks to reduce the routing wirelength and widen FPGA employability for low-power applications with non-volatile property. Among many technologies, we focus on Oxide Resistive Memory (OxRRAM) and Conductive Bridge Resistive Memory (CBRAM) devices by assessing unique properties of these technologies in circuit design. As another solution, we design a new FPGA with 3D monolithic integration (3DMI) by utilizing high-density interconnects. Starting from two layers with logic-on-memory approach, we examine various partitioning schemes with increased number of integrated active layers to reduce the routing complexity and increase logic density. Based on the obtained results, we demonstrate that multi-tier 3DMI is a strong alternative for future scaling.
PhD degrees awarded in 2014

Yassine Fkih
University: Université Montpellier II

Design For Test of TSV based 3D Stacked Integrated Circuits

For several years, the complexity of integrated circuits continues to increase, from SOC (System On Chip) to SIP (System In Package), and more recently 3D SiCs (Stacked Integrated Circuits) based on TSVs (Through Silicon Vias) that vertically interconnect stacked circuits in a 3D system. 3D SiCs have many advantages in terms of small form factor, high performances and low power consumption but have many challenges regarding their test which is a necessary step before the commissioning of these complex systems. In this thesis we focus on defining the test infrastructure that will detect any occurring defects during the manufacturing process of TSVs or the different sacked chips in the system. We propose a BIST (Built In Self Test) solution for TSVs testing before stacking, this solution is based on the use of ring oscillators, whose oscillation frequencies depend on the electrical characteristics of the TSVs. The proposed test solution not only allows the detection of faulty TSVs but also gives information about the number of defective TSVs and their location. On the other hand, we propose a 3D DFT (Design For Test) architecture based on the new proposed test standard IEEE P1687. The proposed test architecture provides test access to the components of the 3D system before and after stacking. Also it allows the re-use of recycled test data developed and applied before stacking to each die in the mid-bond and post-bond test levels. This work lead to the opening of a new problem related to the test scheduling under constraints such as: power consumption, temperature.

Yeter Akgul
University: Université Montpellier II

Power Management based on Dynamic Voltage, Frequency and Body Bias Scaling on System On Chip in FD-SOI technology

Beyond 28nm CMOS BULK technology node, some limits have been reached in terms of performance improvements. This is mainly due to the increasing power consumption. This is one of the reasons why new technologies have been developed, including those based on Silicon-On-Insulator (SOI). Moreover, the standardization of complex architectures such as multi-core architectures emphasizes the problem of power management at fine grain. FD-SOI technologies offer new power management opportunities by adjusting, in addition to the usual parameters such as supply voltage and clock frequency, the body bias voltage. In this context, this work explores new opportunities and searches novel solutions for dynamically manage supply voltage, clock frequency and body bias voltage in order to optimize the power consumption of System on Chip. Adjusting supply voltage, frequency and body bias parameters allows multiple operating points, which must satisfy the constraints of functionality and performance. This work focuses initially at design time, proposing a method to optimize the placement of these operating points. An analytical solution to maximize power savings achieved through the use of several operating points is provided. The second important contribution of this work is a method based on convexity concept to dynamically manage the supply voltage, the frequency and the body bias voltage so as to optimize the energy efficiency. The experimental results based on real circuits show average power savings reaching 35%.
PhD degrees awarded in 2014

Victor Lomuller
University: Université Grenoble Alpes

Multi-Time Code Generation and Multi-Objective Code Optimisation

Compilation is an essential step to create efficient applications. This step allows the use of high-level and target independent languages while maintaining good performances. However, many obstacles prevent compilers to fully optimize applications. For static compilers, the major obstacle is the poor knowledge of the execution context, particularly knowledge on the architecture and data. This knowledge is progressively known during the application life cycle. Compilers progressively integrated dynamic code generation techniques to be able to use this knowledge. However, those techniques usually focus on improvement of hardware capabilities usage but don’t take data into account. In this thesis, we investigate data usage in applications optimization process on Nvidia GPU. We present a method that uses different moments in the application life cycle to create adaptive libraries able to take into account data size. Those libraries can therefore provide more adapted kernels. With the GEMM algorithm, the method is able to provide gains up to 100% while avoiding code size explosion. The thesis also investigates runtime code generation gains and costs from the execution speed, memory footprint and energy consumption point of view. We present and study 2 light-weight runtime code generation approaches that can specialize code. We show that those 2 approaches can obtain comparable, and even superior, gains compared to LLVM but at a lower cost.

Nicolas Hili
University: Université Grenoble Alpes

A Method for Embedded Systems Collaborative Development

Embedded system development is complex and this complexity has several sources. A first one is embedded system own specificities (physical world measurement and control, execution on a physical resource-constrained platform, reliability, responsiveness,…) that distinguish themselves from software systems. Another one comes from industrial concerns about which these systems are subject to product and development costs and delays, multidisciplinary teams, system documentation and certification. To handle this complexity, few methods and languages have been proposed. They focus on a modeling of both application and platform part included in an embedded system. However, the notions of method and process model are barely known from the embedded system community and current methods do not capitalize on the knowledge acquired by other engineering domains like information systems. The goal of this thesis is the definition, the formalization and the tooling of an embedded system development method. To do that, this work focuses on four main contributions : (1) the formalization of a guided process and a language to ensure a consistent modeling of both the application and the platform, (2) the composition of complex platforms to permit a progressive implementation of an application on its concrete platform, (3) the integration of a project management and a product traceability allowing the project manager to measure and monitor the development progress, to organize his team and to parallelize the development, and (4) the development of a tool designed to support the process, the language and the project management.
Continuous deployment of pervasive applications in dynamic environments

Driven by the emergence of new computing environments, dynamically evolving software systems makes it impossible for developers to deploy software with human-centric processes. Instead, there is an increasing need for automation tools that continuously deploy software into execution, in order to push updates or adapt existing software regarding contextual and business changes. Existing solutions fall short on providing fault tolerant, reproducible deployments that would scale on heterogeneous environments. This thesis focuses on enabling continuous deployment solutions for dynamic environments, such as Pervasive Computing environments. It adopts an approach based on a transactional, idempotent process for coordinating deployment actions. This thesis proposes a set of deployment tools, including a deployment manager capable of conducting deployments and continuously adapting applications according to the changes in the current state of the target platform. The implementation of these tools, Rondo, also allows developers and administrators to code application deployments thanks to a deployment descriptor DSL. Using the implementation of Rondo, the propositions of this thesis are validated in several industrial and academic projects by provisioning frameworks as well as installing applications and continuous reconfigurations.

Compilation of Parametric Dataflows Applications for Software-Defined-Radio-Dedicated MPSoCs

The emergence of software-defined radio follows the rapidly evolving telecommunication domain. The requirements in both performance and dynamicity have engendered software-defined-radio-dedicated MPSoCs. Specialization of these MPSoCs makes them difficult to program and verify. Dataflow models of computation have been suggested as a way to mitigate this complexity. Moreover, the need for flexible yet verifiable models has led to the development of new parametric dataflow models. In this thesis, I study the compilation of parametric dataflow applications targeting software-defined-radio platforms. After a hardware and software state of the art in this field, I propose a new refinement of dataflow scheduling, and outline its application to buffer size's verification. Then, I introduce a new high-level format to define dataflow actors and graph, with the associated compilation flow. I apply these concepts to optimized code generation for the Magali software-defined-radio platform. Compilations of parts of the LTE protocol are used to evaluate the performances of the proposed compilation flow.
Greetings

Editorial Committee
Marc Belleville
Christian Gamrat
Hélène Vatouyas
Jean-Baptiste David

Readers
Thierry Collette

Graphic Art
Hélène Vatouyas

Front page illustration copyrights
iStockphoto & J.Gallon
CEA
Pixabay-CC0 Public Domain
Architecture and IC Design, Embedded Software

Contacts

Thierry Collette
Head of Architecture, IC Design & Embedded Software division
thierry.collette@cea.fr

Marc Belleville
Research director, Scientific director-Grenoble
marc.belleville@cea.fr

Christian Gamrat
Research director, Scientific director-Saclay
christian.gamrat@cea.fr

Marc Duranton
Research director, European Projects Coordinator
marc.duranton@cea.fr

Paul Greve
Sr Business Development Manager
paul.greve@cea.fr

Michel Durr
Sr Business Development Manager, Analog & Wireless IC
michel.durr@cea.fr

Denis Dutoit
Strategic Marketing Manager, Computing
denis.dutoit@cea.fr

Bénédicte Robin
Sr Business Development Manager, Digital IC
benedicte.robin@cea.fr

Frédéric Surleau
Sr Business Development Manager, Embedded Systems
frедерюc.суrлеаu@сea.fr

Michael Tchagopian
Sr Business Development Manager, Imaging & Smart Sensors
michael.tchagopian@cea.fr