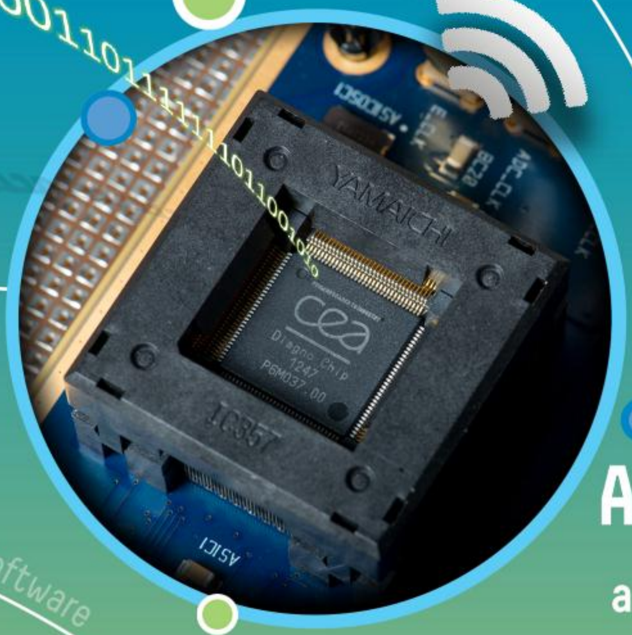
leti list innovation for industry

Annual Research Report 2013



Architecture and IC Design, Embedded Software

DE LA RECHERCHE À L'INDUSTRIE
cea

Annual Research Report 2013 • Leti List

Architecture and IC Design, Embedded Software

Design, Architecture & Embedded Software Division

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 8,000m² 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 1,700 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 650 researchers, engineers and technicians, the CEA-LIST performs innovative research in partnership with major industrial players in the fields of ICT, Energy, Transport, Security & 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 280 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.

Annual Research Report 2013 • leti list

Architecture and IC Design, Embedded Software

Contents

Architecture and IC Design, Embedded Software

Edito

Thierry Collette
Head of the Architecture, Design and Embedded Software Division

More than ever, considering the wave of the Internet of Things / Cloud of Things, the constraints of software / hardware integration within integrated and embedded systems are a priority. Indeed, the IoT, as an extension of the embedded systems domain, is a major and wide issue, that addresses the devices, servers and services.

For these three areas, system and component integration is the key, involving major challenges as ultra-low power (in sensors and actuators, communications, computing from smart devices to servers, and energy harvesting, conversion and management), easy and scalable deployment, reliability, dependability, security and privacy in these new fully connected infrastructure.

All these issues are driving our current research activities, aiming to provide tomorrow innovative solutions to our industrial partners. These topics are addressed in this report with sections dedicated to Smart Interconnected Devices, Digital Architectures & Systems, Power & Temperature Optimized Digital Circuits, and Circuits in Emerging Technologies.

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

Thierry Collette

Architecture and IC Design, Embedded Software

Annual Research Report 2013 • leti list

Architecture and IC Design,
Embedded Software

Key Figures

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

174 Permanent researchers,
65 PhDs and Post-docs

Design & Embedded system plateform
Full suite of IC CAD and Embedded system tools,
Hardware Emulators,
& Test equipments,

34M€ budget
85% funding from contracts

orquati · Koen Bertels
Karlsson · François Pacull Edito
Springer
Smart
Multicore
Embedded
ystems

46 granted patents
38 papers, journals & books
175 conferences & workshops

Architecture and IC Design, Embedded Software

Scientific Activity

Publications

175 publications in 2013, including journals and Top conferences like ISSCC, VLSI Circuits Symposium, ESSCIRC, CICC, IMS, ISCAS, DAC, DATE, ACC, ECC, RTSS, IPDPS and ESWeek

Distinctions and Awards

Catrene Innovation Award – Panama project

Nanoelectronic Forum 2013 Exhibition Awards: 2nd place SEEL project

IEEE 3PGCIC'13 - Best Paper Award – Oana Stan

4th European Workshop on CMOS variability – Best Student Paper Award – Lionel Vincent

RTNS 2013 – Best Student Paper Award – Vincent Legout

Expertises and Recognitions

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

Scientific Committees

Editorial Boards: IEEE TCAS I, Journal of Low Power Electronics,

19 members of Technical Programs and Steering Committees in major conferences: ISSCC, ESSCIRC, DAC, DATE, ASP-DAC, ESWEEK, RTNS, IJCNN, IWANN, EMSOFT, NANOARCH…

Normalization committee: AUTOSAR (Automotive Open System Architecture)

International Collaborations

Collaborations with more than 20 universities and institutes worldwide
Caltech, University of Berkeley, University of Columbia, Carnegie Mellon University, EPFL, CSEM, UCL, Polito Torino, KIT, Chalmers University, Tongji, ….

Architecture and IC Design, Embedded Software

1

Smart Interconnected Devices

Imagers, Image Processors

Sensor Interfaces

Energy conversion

RF communications

Localization

Middleware for IoT

Diagnostic

A novel 0.5GHz real time single-photon detection technique: circuit design for cooled HgCdTe infrared APD detector

Research topics : Single-photon detection, photon counting, infrared sensor

H. Amhaz, K. Foubert , F. Guellec, and J. Rothman

ABSTRACT: A readout IC dedicated to mono-element photon detection in the SWIR infrared band has been developed. Its multi-channel architecture enables high detection rate. Each channel is hybridized to a HgCdTe avalanche photodiode (APD) that provides gain with low excess-noise. The input-referred channel noise is 12 electrons allowing single-photon detection with a APD gain (M) in the 30 to 50 range. For M=50, the detection efficiency reaches 98% with a 5ns p-p precision in the detection when taking into account the channel-to-channel mismatch. The channel detection rate is about 32MHz leading to a global rate above 0.5GHz.

The HgCdTe Avalanche Photodiode (APD) technology initially developed at CEA-Léti for infrared imaging application is also interesting for several mono-element applications due to the high gain, high bandwidth and low excess noise characteristics of this device.
Atmospheric and telemetric LIDAR applications require a high bandwidth (20MHz to 100MHz) and high sensitivity. Free-space telecommunication will require higher bandwidth (100MHz to several GHz). Other applications like spectrometry and photoluminescence require a very high sensitivity in order to detect a single photon.

In this context, we developed a versatile circuit that could satisfy various needs [1]. The main applications require a minimum photosensitive area of 100x100µm². From a technological point of view it is difficult to make large APD with good performances. Consequently, a 5x5 diode array has been used to cover the surface based on a 25µm pitch.
The read-out IC (ROIC) specifications needed to cover a large panel of applications are quite demanding. In addition to real-time single-photon detection the mono-element detector should be available to work over a high input photon rate range.
A multi-channel approach has been adopted in order to increase the global detection rate (Fig. 1). Each APD is connected to an event detector and the digital outputs of the channels are added to provide a 5-bit asynchronous data representing the events detected over the whole area.

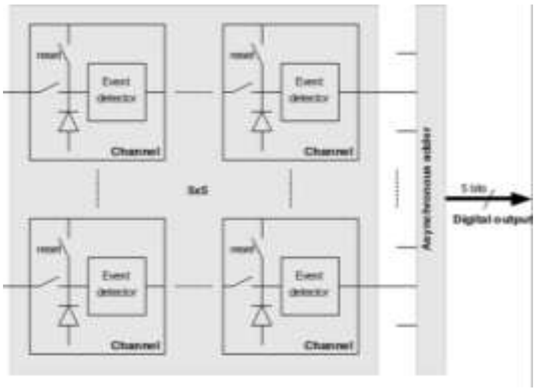


Figure 1: Multi-channel detector architecture

The input current is directly integrated on the photodiode junction capacitance producing a staircase signal with small steps (250µV per photon for an APD gain M=30). This signal is amplified and filtering is then applied to use a detection scheme based on transition rather than level (Fig. 2).

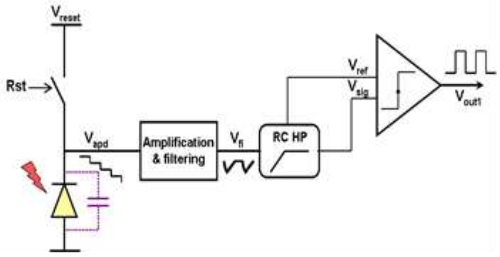


Figure 2: Bloc diagram of the event detection circuit

The simulated input-referred temporal noise is about 12 electrons. This performance allows single-photon detection with low false count rate (10e-3 to 10e-5) at moderate APD gain (30 to 50). The event detection circuit works asynchronously and is able to distinguish 2 photons separated by 31ns. Thanks to the detector's asynchronous multi-channel architecture, the 32MHz channel rate lead to a global photon detection rate above 0.5GHz when considering uniform illumination over a 100x100µm² area (4x4 channels with 25µm pitch).

In the chosen architecture, channel-to-channel mismatch is also a key issue that may limit the temporal precision of the detection. The circuit has been designed to have a low sensibility to amplifier and comparator offsets. As a result the channel-to-channel delay variation is around 5ns p-p. Mismatch also affects Photon Detection Efficiency (PDE). Monte-Carlo simulations have shown that a PDE of 98% is obtained for an APD gain of 50. While it is not possible to perform single-photon detection with M<20, the PDE is still rather good at 85% for M=30.

The circuit has been fabricated in a standard 0.18µm 1.8V/3.3V CMOS process. The goal is to demonstrate the advantage of this detector technology by implementing the prototype in different experimental setups covering several applications.

Related Publications :
[1] H Amhaz, K. Foubert, F. Guellec, et al., "A Novel 0.5GHz Real Time Asynchronous Photon Detection and Counting Technique: ROIC Design for Cooled SWIR HgCdTe Infrared Detector", IEEE International NEWCAS conference, 2013.

A [10°C ; 70°C] 640×480 17µm Pixel Pitch TEC-Less IR Bolometer Imager with Below 50mK and Below 4V Power Supply

Research topics : Thermo Electrical Cooler Less, Infra Red, Imager, Bolometer

B. Dupont, A. Dupret, S. Becker, A. Hamelin, F. Guellec, P. Imperinetti, W. Rabaud

ABSTRACT: Used in low cost thermal imaging, infrared micro-bolometer detectors are very demanding in terms of offset skimming and technological fluctuation compensation. To reach a noise equivalent temperature difference (NETD) about 50mk, and considering the dependence of both offset and Fixed-Pattern Noise (FPN) on the temperature of the focal plane, a thermo electrical cooler (TEC) is used to prevent the temperature of the focal plane from varying either spatially or temporally. This paper presents an architecture that allows to get rid off the TEC, while keeping performances at the state of art.

This circuit is based on an enhanced differential pixel read out (Fig.1.). As all micro-bolometer imagers, this circuit is based on sensitive thermistor arrays. Bolometer resistance varies according to its temperature, which is linked to biasing, self-heating, focal plane temperature and scene temperature. The difference between the sensitive bolometers and a reference bolometer gives an image of the scene temperature. This difference is less accurate when the characteristics of the reference bolometer differ from the sensitive bolometer (polarization, duty cycle and physical design). The read out circuit also impacts the accuracy of the difference and its sensitivity to temperature. To sum-up the TEC-less characteristic implies to keep output NETD over a large thermal range without tuning supplies or polarizations of the circuit. So it depends on conversion gain, mismatch and offsets level drift.

In the presented circuit, the reference is provided by a column of shielded bolometer with the same characteristics as the sensing bolometer and located at the head of rows of the pixel array. Since the pixel array is read in a rolling shutter way, the shielded bolometer has the same thermal cycle as the sensing bolometers. For both the reference and the sensing branch, PMOS voltage followers with identical gate voltage force identical voltage across both bolometers.

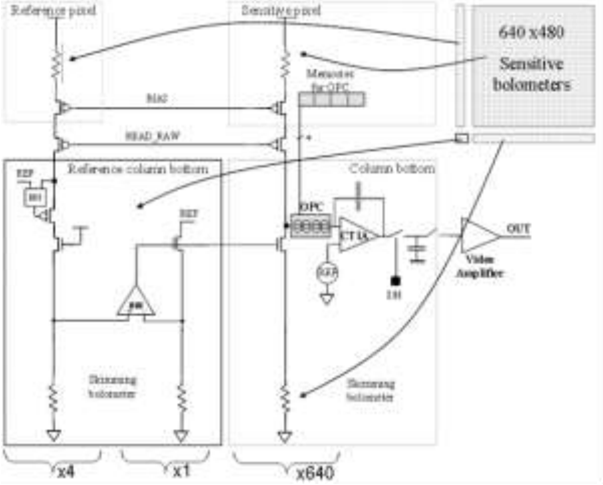


Figure 1: Differential pixel readout architecture

A Buffered Direct Injection (BDI) forces Vref on the drain of the PMOS of the reference branch. It prevents Early effect and allows a more accurate copy of the current. To keep the reference branch noise below that of the sensing branch, dominated by 1/f noise, large NMOS are designed. As a result the BDI load about 800pF, and rejects the 1/f noise by a factor of 50 in the [1Hz ; 5KHz] bandwidth.

Regarding the transimpedance amplifier (CTIA), its gain can be set between 5 and 70V/µA. At 30°C such a high gain allows having an output NETD only 2% higher than the intrinsic sensing bolometer NETD.

To compensate the remaining FPN, an offset pre-correction (OPC) is implemented: thermalized bolometers provide currents {Ioff, Ioff/2, Ioff/4, Ioff/8} that are combined according to the 4-bit values stored in each pixel. By reducing mismatch impact before integration OPC prevents the CTIA from saturating under thermal drift, especially at high gain. By being less sensitive to thermal drift, a low gain would enable a larger TEC less range but the read out chain would intolerably degrade the output NETD.

Conversely for a given TEC-less range, the OPC enables reducing the constraints on bolometer mismatches. This constitutes a key factor when low-cost micro-bolometer imagers are targeted.

The readout chain accounts for about 95% of the 170mW total power consumption.

Designed in 0.18µm CMOS technology presented ROIC allows a TEC-less capability with a low 4V power supply without using off-chip compensation tables. To the best of our knowledge, this micro-bolometer imager is the only 17µm pixel pitch that features a NETD less than 50mK over a [10°C ; 70°C] operating range.

The differential architecture is naturally TEC-less whatever the bolometer resistance values are.

The digital OPC reduces the impact of FPN and contribute to extend TEC-less range. Finally the enhanced differential architecture also rejects supply noise by a factor of 50, making this circuit fully compliant with harsh environment.

Related Publications :
[1] Bertrand Dupont, Antoine Dupret, Sebastien Becker, Antoine Hamelin, Fabrice Guellec, Pierre Imperinetti, Wilfried Rabaud: A [10°C; 70°C] 640×480 17µm pixel pitch TEC-less IR bolometer imager with below 50mK and below 4V power supply. ISSCC 2013: 394-395

A 120µW 240×110@25fps vision chip with ROI detection SIMD processing unit

Research topics : CMOS image sensors, low power, analog processing

A. Verdant, A. Dupret, P. Villard, L. Alacoque, H. Mathias (IEF), F. Delgehier (IEF)

ABSTRACT: A smart ultra-low power CMOS image sensor comprising an analog programmable processor array is reported. Compact and efficient motion detection algorithms are implemented to process sub-sampled images made of so-called macropixels. Only Regions of Interest (ROI) consisting of macropixels containing moving objects are read out. This drastically reduces power consumption: the 110×240 pixel image sensor fabricated in a 0.35µm technology features a power consumption of 120µW at 25fps.

Achieving ultra-low power consumption is of most importance for battery powered image sensors. In the case of video surveillance applications, this power consumption can be adapted to the scene activity. However, the detection of intrusions implemented on processors associated to video cameras wastes a considerable amount of power. Indeed, every frame of the whole image array is processed, whereas the actual scene is steady, or merely comprises very few and small regions of interest (ROIs).

A CMOS image sensor capable of efficiently detecting and tracking ROIs with ultra-low power consumption has been designed. The circuit has been optimized to implement various motion detection algorithms, from which derives the locations of ROIs. Those algorithms based on the difference between the actual scene and an estimate of its background, obtained from the temporal filtering of the macropixels' luminance, are among the most compact ones [1]. The image sensor proof of concept is composed of a 240×110 pixel array associated to a Single Instruction Multiple Data (SIMD) vector of 11 general purpose analog programmable processing units (PUs) and to 5 banks of 24×11 analog memory array (ARAM) (Figure 1).

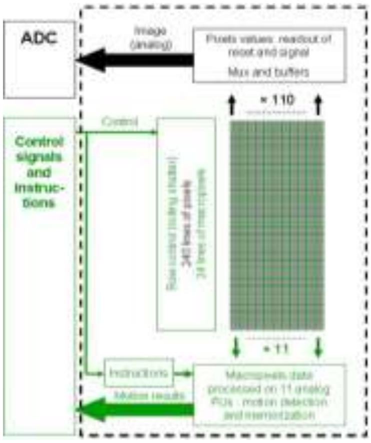

ADC
Image (analog)
Pixels values: readout of reset and signal
Mux and buffers
× 110
Control
Control signals and instructions
Row control (rolling shutter)
240 lines of pixels
24 lines of macropixels
× 11
Instructions
Motion results
Macropixels data processed on 11 analog PUs : motion detection and memorization

Figure 1 : Global ASIC architecture

The computations are performed on the local average of 10×10 pixels. These averages are acquired in rolling shutter mode, as well as full resolution sub-images (ROIs).

By using this sub-sampling scheme, the data throughput towards the processing elements is divided by 100, while keeping the contribution of all the pixels.

The results of the computations are used to detect critical temporal activity, considered as moving objects, and to give the location of the regions of interest (ROIs) where moving objects have been located. Only the pixels within the macropixels elected as ROIs are driven outside the sensor by a readout pipeline opposite to the vector of PUs. As long as no motion appears, the high resolution pixels are not read out, which contributes to further reduce the power consumption. The general sensor behavior is exposed on Figure 2.

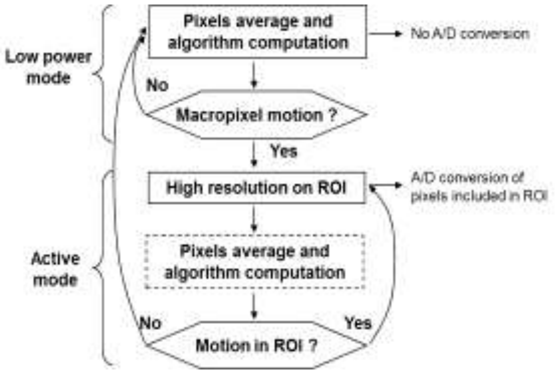

Low power mode
Active mode
Pixels average and algorithm computation
No A/D conversion
No
Macropixel motion ?
Yes
High resolution on ROI
A/D conversion of pixels included in ROI
Pixels average and algorithm computation
No
Yes
Motion in ROI ?

Figure 2 : Synoptic describing sensor behavior for motion detection processing and high resolution pixel readout

Various compact motion detection algorithms have been implemented, achieving Detection Rate as high as 95%.An image acquired with the presented sensor is exposed on Figure 3.

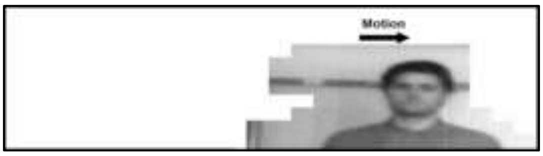


Figure 3 : Example of ROI detection

The total power consumption of the pixel array, PUs and memory is 120µW. Using standard batteries of about 1 A.h, the autonomy of the presented sensor reaches one year.

Related Publications :

[1] Arnaud Verdant, Patrick Villard, Antoine Dupret, and Hervé Mathias, "Three Novell Analog-Domain Algorithms for Motion Detection in Video Surveillance," EURASIP Journal on Image and Video Processing, vol. 2011, Article ID 698914, 13 pages, 2011. doi:10.1155/2011/698914.

[2] Arnaud Verdant, Antoine Dupret, Patrick Villard, Laurent Alacoque, Hervé Mathias, and Flavien Delgehier, "A 120µW 240×110@25fps vision chip with ROI detection SIMD processing unit", ISCAS, page 2412-2415. IEEE, (2013)

3D integrated burst CMOS image sensor for high speed applications

Research topics: High Speed Imaging, Burst Image Sensor, 3D Integration

R. Bonnard, F. Guellec, J. Segura, A. Dupret, W. Uhring (ICube)

ABSTRACT: As 3D integration appears to be a key enabling technology for future image sensors, it is especially true for high speed imaging. We present two new 3D stacked architectures for burst high speed imaging with in-situ A/D conversion, one with analog memories, the other based on digital storage. Thanks to our research, analog storage architecture appears as a good way to reach unmatched speed performances (>10 Terapixel/s) while digital storage architecture seems an interesting solution to reach high memory depth (>200 frames).

High speed imaging is a cutting edge technology useful in many technical and scientific fields as a measuring instrument and a monitoring tool. These sensors are used to study fast phenomena like fracture mechanics, fluorescence life time, plasma forming, etc. Two kinds of high speed image sensor (IS) must be distinguished: continuous IS and burst IS. In continuous mode, the image sensor acquires and reads out of the chip the frames one after the other. However due to the limited number of output channels, the readout circuits are the bottlenecks which limit the frame rate to tens of kilo-frames-per-second (fps). To overcome this limit, a solution is to acquire at higher rates and store a burst of frames on chip before read out. 3D integration appears to be a key technology to design burst image sensor. Indeed it allows to perform highly parallel acquisition and to store a large number of frames while keeping a good pixel fill factor (i.e. good sensitivity).

We propose two architectures of burst image sensor with on-chip A/D conversion, one with analog storage, Fig. 1 and one with digital storage, Fig. 2. For the analog storage architecture, the front end electronic captures (1st tier) and stores (2nd tier) the burst in an analog memory. Then, the ADCs convert (3rd tier) the frames into digital data which are serialized read out through the output driver (3rd tier).

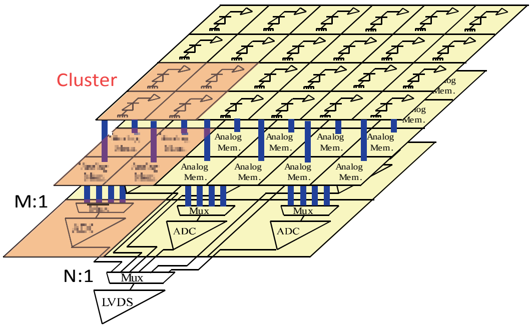


Figure 1: Burst image sensor with analog storage architecture

In digital storage architecture the A/D conversion is performed before the frame storage as shown. First, the pixel signals (1st tier) are converted into digital signals (2nd tier) which are stored into digital memories (3rd tier). Then, the digital memory is read out of the pixel through an output driver (3rd tier).

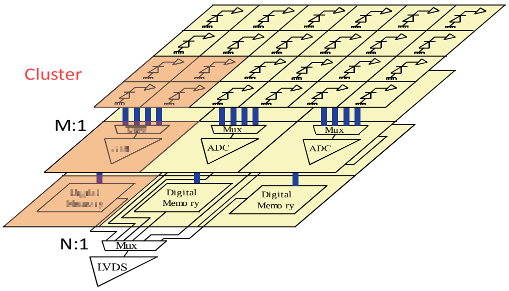


Figure 2: Burst image sensor with digital storage architecture

In [1] we propose appraisals of frame-rate, memory depth and power consumption for both architectures which are summarized in Fig. 3.

Storage	Pixel Size (µm)	Resolution (pixel)	Frame Rate (Mfps)	Memory Depth (frame)	Power (W)
Analog	40x40	500x500	100-1000	~80	2
Digital	40x40	500x500	1-10	>200	11

Figure 3: Performance assessment

The architecture analysis shows that digital storage architecture has a slower frame rate (up to 10 mega-fps) than analog storage architecture but has a higher memory depth which can easily exceed 200 frames for a 40µm pixel pitch. In terms of power consumption, the analog storage architecture is more energy efficient than the digital one. Moreover during the burst acquisition, the power consumption of both architectures skyrockets due to memory accesses. Compared to the state of the art (SOA), on one hand, 3D analog storage architecture would allow 100 to 1000 mega-fps while SOA reaches 10-20 mega-fps for 2D image sensor with about the same memory depth. On the other hand, 3D digital storage architecture allows to store more than 200 frames per burst while present burst image sensors do not exceed 150 frames.

Related publication:
[1] Bonnard, R.; Guellec, F.; Segura, J.; Dupret, A.; Uhring, W., "New 3D-integrated burst image sensor architectures with in-situ A/D conversion," 2013 Conference on Design and Architectures for Signal and Image Processing (DASIP), vol., no., pp.215-222, 8-10 Oct. 2013

Co-integration of a Smart CMOS Image Sensor and a Spatial Light Modulator for Real-Time Optical Phase Modulation

Research topics: CMOS Image Sensor, Optical phase modulation, Acousto-optic

T.Laforest, A.Verdant, A. Dupret, S. Gigan*, F. Ramaz*, G. Tessier* (*ESPCI ParisTech)

ABSTRACT: we present a CMOS image sensor architecture coupling a spatial light modulator to a special scheme photodiode, for medical imaging based on acousto-optical coherence tomography with a digital holographic detection. Our 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 16 µm pixel pitch, at 4000 Hz. This is compliant with correlation time of breast tissues. The stacking of a photosensitive element with a spatial light modulator on the same device brings a significantly higher robustness compared to the state of the art techniques.

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 [2]. Indeed, such a technique enables foreseeing great progress in breast medical imaging in the near future.

However, its clinical application is still out of reach because of the complexity of the setup and the limitations of the detection scheme. One of the major problems is that in thick biological tissues, the correlation time of the transmitted intensity through the sample is typically a few milliseconds. Furthermore, the relevant signal (i.e. useful to extract optical information) corresponds to 10% of the total incident light power on the sensor. As a consequence the detector must have at the same time a frame rate higher than 2 kfps and a noise level compliant with the detection of the scattered beam.

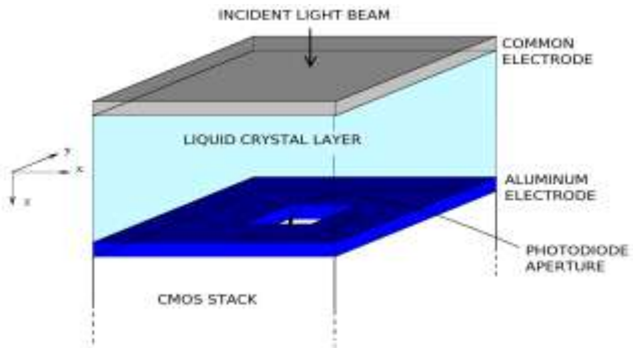


Figure 1 : Circuit for voltage gain calculation

Moreover, coupling acousto-optic holographic scheme with a spatial phase modulation setup allows phase conjugation and light focusing through the sample on a region of interest. This focusing would make detection much easier, e.g. by improving the acousto-optic signal coming from the region of interest. State of the art setups are based on the association of a camera, which acts as a wavefront sensor, and a spatial light modulator. The camera sends the measured phase of the wavefront to a processing unit, which in turn sends a feedback control to the modulator for phase-conjugation and focusing of light on the regions of interest. Such a principle has been recently investigated using different techniques and shows a great interest not only for imaging diagnosis but also for therapy in thick biological media. Nevertheless, the phase conjugation setup presents several limitations. Firstly, the modulator and the camera pixel arrays must be perfectly matched spatially but this is at the cost of complicated mechanical processes, both bulky, complicated to align and expensive. Secondly, separated elements lead to a feedback delay that prohibits controlling light, corresponding typically to some hundreds of milliseconds, which is incompatible with the analysis of living tissues.

In order to address these limitations, we propose an architecture for a CMOS image sensor that is dedicated to wavefront acquisition and modulation. The pixel integrates a photodiode with some analog primitives and a spatial light modulator (SLM) made of liquid crystal covering the entire pixel. This stacking, shown in Fig 1, allows a perfect matching between photodiode and SLM pixels. The pixel to SLM connectivity circumvents the latency due to rolling mode readout. As a consequence this stack allows a delay free feedback between photodiode data to SLM control, making the image sensor compliant with biological media correlation time constrains.

An architecture for an original pixel coupling a spatial light modulator and an image sensor has been presented in [1,2]. Optical simulations have also been presented in [3].

Related Publications:
[1] Laforest, T.; Dupret, A.; Verdant, A.; Ramaz, F.; Gigan, S.; Tessier, G.; Benoit a la Guillaume, E., "A 4000 Hz CMOS image sensor with in-pixel processing for light measurement and modulation," New Circuits and Systems Conference (NEWCAS), 2013 IEEE 11th International , vol., no., pp.1,4, 16-19 June 2013 .
[2] Laforest, T.; Dupret, A.; Verdant, A.; Ramaz, F.; Gigan, S.; Tessier, G "Co-Integration of a Smart CMOS Image Sensor and a Spatial Light Modulator for Real-Time Optical Phase Modulation.", SPIE Electronic Imaging 2014.
[3] Laforest, T.; Dupret, A.; Verdant, A.; Ramaz, F.; Gigan, S.; Tessier, G," Monolithic device for on-chip fast Optical Phase Conjugation integrating an image sensor and a spatial light modulator", SPIE Photonics West 2014.

New applications and sensing strategies for compressive sensing

Research topics: High Dynamic Range, Compressive Sensing, Features extraction

W.Guicquero, A.Dupret, P.Vandergheynst (EPFL)

ABSTRACT: With the rise of compressive sensing image sensors, it becomes necessary to define relevant sensing schemes and applications. Our work aims at proposing new applications and sensing strategies for that purpose. Since new sensing schemes become practical on real image sensors, side applications will emerge. In this work, we have proposed two novel applications, Compressive Feature Imaging [1] and High Dynamic range Compressive Sensing [2].

The motivation of this work is to generalize the Compressive Sensing (CS) paradigm to novel applications and sensing methods. First we propose a nonlinear sensing strategy providing relevant features [1] for classification purpose -for example- and secondly a High Dynamic Range CS multi-capture method [2].

We demonstrate in [1] that block variance is a good candidate for tuning the number of measurements performed by block in the context of adaptive block-based CS. It is also shown that block variances can be used as nonlinear CS measurements. In addition, such a statistical descriptor could easily be implemented in the analog domain before the analog to digital conversion into an image sensor. Those features can also be used as features for tracking or classification purposes. The best reconstruction results shown in figure 1 are obtained using a proposed constraint operator on the image x. Op(x) is composed of multiple wavelet transforms (mWT) on horizontal and vertical gradients (TV).

$$Op(x) = \|mWT(TV(x))\|_1 \qquad \text{Eq.1}$$

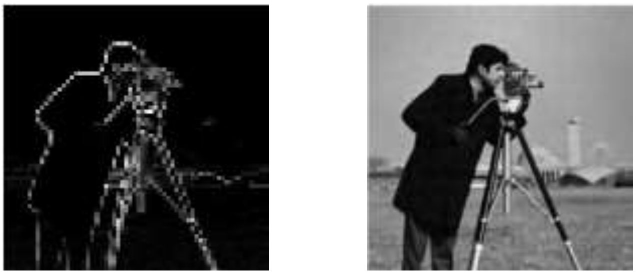


Figure 1: Local variance measurements and the reconstructed image (PSNR = 31.1 dB) with a compression ratio of 7.8%.

Multi-capture HDR implies to store full frame images. This technique is based on acquiring multiple images from a single scene using multiple different times of exposure. The goal is thus to acquire low illumination regions with a long time of exposure and respectively high illumination regions using a short time of exposure. This way, a HDR image can be reconstructed from those different images preserving low dynamic details in low illumination regions and eliminating saturation problems for highest illumination regions.

However, it requires either a lot of memory resources or digital processing power to perform compression. CS can reduce a lot those memory requirements at the sensor level. Assuming that a HDR image is sparse in some sense, the image can be reconstructed from relatively few measurements performed on images corresponding to different times of exposure. In the proposed method, a global tone-mapping (tm) has been added to make the reconstructed image of a better quality in terms of visual rendering. The proposed constraint operator on RGB image x is expressed in Eq.2. It is composed of multiple wavelet transforms (mWT) on the YUV color representation of the tone mapped version of x joint with the Total Variation (TV) of tm(x). Figure 2 shows an example of an HDR CS reconstruction.

$$Op(x) = \|mWT(YUV(tm(x))) + TV(tm(x))\|_1 \qquad \text{Eq.2}$$

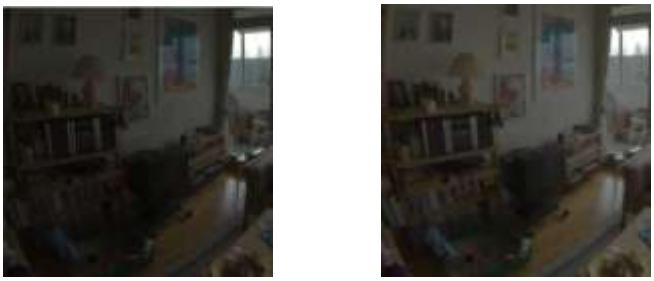


Figure 2: Original16-bit image and its tone mapped reconstruction (PSNR = 42 dB) (demosaicing, dynamic range increasing of 18dB).

This work exposes a novel approach for multi-capture imaging taking advantage of the CS paradigm. The proposed technique performs at the same time a HDR reconstruction for the three color channels. It opens a new way to design compressive sensing image sensor particularly adapted to acquire HDR images. The dynamic range of the acquisition can be largely increased without imposing a huge amount of data comparing to a classical multi-capture. In addition, a new set of constraints is proposed for the color image reconstruction problem, referring to demosaicing. Those constraints are applied on different color spaces to jointly reconstruct color channels of the demosaiced HDR image.

Related Publications:
[1] W. Guicquero, A. Dupret, P. Vandergheynst, "A adaptive compressing sensing with side information" (IEEE ASILOMAR CSS&C), 2013.
[2]W. Guicquero, A. Dupret, P. Vandergheynst, "Multi-capture high dynamic range compressive imaging" (IEEE ASILOMAR CSS&C), 2013.

Hardware Implementation for Entropy Coding and Byte Stream Packing Engine in H.264/AVC

Research topics : H.264/AVC, entropy coding, CAVLC, Exp-Golomb, Video byte stream

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

ABSTRACT: Entropy coding (EC) and data packing are the major phases in video coding. The H.264 Advanced Video Coding (H.264/AVC) standard adopts Exp-Golomb and Context-Adaptive coding methods to increase data compression ratio. We propose a hardware architecture of EC and byte stream data packing engines for the H.264/AVC. Our EC engine, containing Exp-Golomb and Context-Adaptive Variable Length Coding (CAVLC), supports the baseline and main profiles of the standard. The proposed architecture is implemented using 180nm technology from AMS. The design consumes only 1.56mW at the operating frequency of 100MHz.

Recommended by both the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG), the H.264 Advanced Video Coding (H.264/AVC) can save approximately 50% of bit rate in comparison with previous standards **Erreur ! Source du renvoi introuvable.**thanks to the adoption of several features. The H.264/AVC also provides high loss/error robustness using separated parameter set structures, which keep key information. For instance, the NAL unit syntax structure enables "network friendliness" to customize the use of the Video Coded Layer (VCL) for various systems and networks.

Many works propose hardware (HW) implementation of the standard. However, they mostly focused on the prediction part of encoders/decoders and on the improvements of algorithms. It can be seen in the literature that for EC in baseline profile (i.e. the Exp-Golomb and CAVLC coding techniques), hardware EC engines usually only encode the data at MB level, i.e. SPS, PPS and Slice header might be implemented in software.

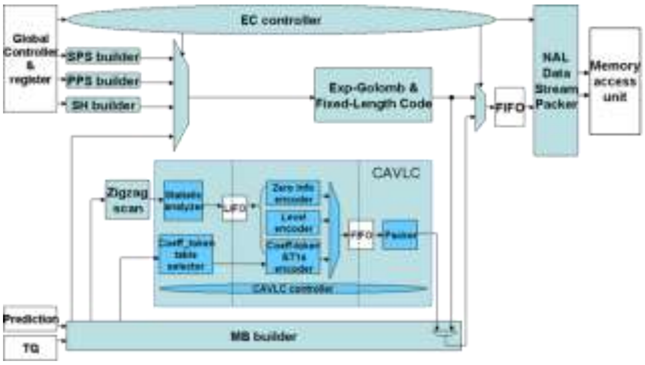


Figure 1: Architecture of entropy coder.

We propose a new architecture of the entropy coding and data packing engine as shown in Figure 1. Encoded syntax elements are transferred from entropy coding to data packing via a FIFO. The main part of the entropy coding engine consists of two encoders, namely, the Exp-Golomb and Fixed-length code (EGF) and CAVLC modules [1]. SPS, PPS, SH, and MB builders are implemented to collect the to-be-encoded data and send them to the EGF module in the specified order. SPS, PPS, SH's information is provided by a global controller via system registers. The MB header information is received from Intra-, Inter- and TQ engines.

The architecture proposed has been modeled in VHDL at RTL level. To verify, our EC and data packing module has been integrated into a hardware H.264 video encoding system. Using the video encoder system with the proposed EC NAL module, we encode raw test video sequences in the CIF format (which is one format we target as it is used for mobile applications). The encoded videos are received from the output of the data packing module. For validation purpose, these encoded videos are then successfully decoded with the JM decoder.

The simulation is done by using ModelSim from Mentor Graphics and the design is synthesized with AMS CMOS 180nm technology by using DC Compiler from Synopsys.

Some implementation results of our EC and byte-stream packing data engines are presented in Table 1.

Table 1: Implementation results of the proposed design

Technology	Cycles/MB	Frequency	Area cost	Power
AMS 0.18µm	691 in the worst case	100MHz	73.5 Kgates	1.56mW at 100MHz

Due to the implementation of the table selector for coeff_token syntax element and the full hardware implementation of the EC, including SPS, PPS, slice header data generators, our design occupies a slightly larger silicon area (approximately 73.5Kgates) than the ones found in the literature.

However, in terms of throughput, our EC engine encodes an MB in maximum 691 cycles (151 cycles for the worst case of MB header plus 540 cycles for the worst case of residual data). With this speed and at the operating frequency of 100MHz, the design is suitable for 720HD video format. Moreover, in average, the encoding process only requires 25 to 90 cycles for MB header and 450 cycles for MB residual.

In terms of power consumption, at 100MHz, it only consumes 1.56mW which is less than most of the low-power designs that operate at 27MHz.

More information on our design can be found in [2].

Related Publications:
[1] N.-M. Nguyen, X.-T. Tran, P. Vivet, and S. Lesecq. An efficient Context Adaptive Variable Length coding architecture for H.264/AVC video encoders. International Conference on Advanced Technologies for Communications (ATC), pp. 158–164, 2012.
[2] N.-M. Nguyen, E. Beigne, S. Lesecq, P. Vivet, D.-H. Bui, X.-T. Tran, Hardware implementation for entropy coding and byte stream packing engine in H.264/AVC, International Conference on Advanced Technologies for Communications (ATC), pp. 360-365, 2013.

GESTURE RECOGNITION ON SMART CAMERAS

Research topics: Gesture Recognition algorithms, embedded systems, Smart camera.

A. Dziri, S. Chevobbe, M. Darouich

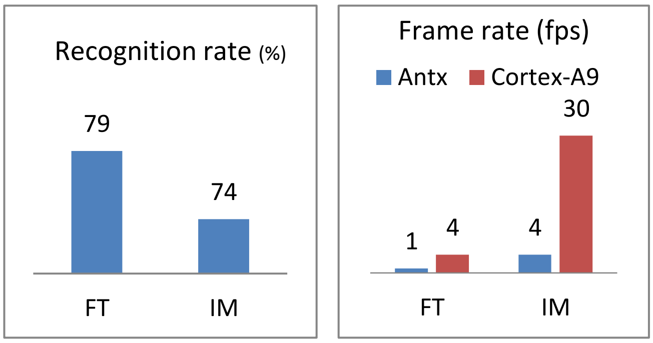
ABSTRACT: Gesture recognition allows a natural interaction without using complex devices. However, most real-time methods are designed to work on computers with high computing resources and memory. This work aims to analyze the relevant methods and investigates the ability of hand gesture recognition on smart camera. Indeed, two gesture pipelines are designed and implemented on embedded processors. The results obtained show that gesture recognition with an acceptable recognition rate (70-80%) can be executed in real-time (4-30 fps) and uses about 200 kB of memory on embedded processors that can be used in smart cameras.

Several gesture recognition methods are developed to allow a natural human-machine interaction. However, these methods are designed without considering the computing resources and the amount of memory available in the embedded systems used in smart cameras. Furthermore, recognition systems use sophisticated sensors like Kinect, Time-of-Flight camera or stereo cameras. Theses sensors are more expensive and less compact than a smart camera. Smart camera is a compact vision system integrating an image sensor, low power embedded processors and low capacity memory on the same chip. This work consists in investigating the gesture recognition for ultra-integrated smart cameras. Gesture recognition on smart camera is a challenging problem because of the low computing resources available: low clock frequency for the processor, no floating point unit (FPU) and a small amount of memory (<1 MB).

To meet the challenge, we proposed two gesture recognition pipelines. Each pipeline is composed of two main blocks as shown in the Fig.1. In the first pipeline, the gesture recognition method is based on the use of invariant moments. The second one is based on finger tips detection. Both pipelines are implemented to comply with the constraints of embedded processors that we find in smart cameras. Indeed, floating-point to fixed-point conversion was performed and memory usage reduced. After, we analyzed for each pipeline its performances (recognition rate, frame rate and memory) on embedded processors.

In this study we targeted two embedded processors Antx [2] and ARM Cortex-A9. Antx is a simple low footprint processor that integrates neither hardware multiplier nor FPU. This processor is close to the embedded processors that we find in the ultra-integrated smart cameras. In the other hand, Cortex-A9 is powerful embedded processor that can be used for high performance smart cameras.

The results of performances analysis of both pipelines are shown in the Fig.2. The recognition rate is given by the Fig.2.a. and the execution time on the embedded processors is given by the frame rate in the Fig.2.b. The memory analysis from our implementation shows that less than 200 kB is required for each pipeline.



a. Recognition rate. *b. frame rate.*

Figure 2: Performances of finger tips and invariant moment pipelines: IM:invariant moments, FT: finger tips.

The results obtained show that the gesture recognition can be executed on embedded processors for few gestures in real-time with an acceptable recognition rate. Future works will consist in implementing the pipeline of finger tips detection on the real smart camera to perform human-machine interaction. Another possible study would be to modify the pipeline to improve the recognition rate.

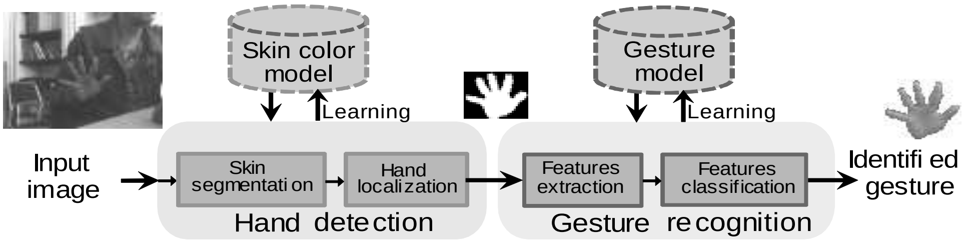


Figure 1: Gesture recognition pipeline

Related Publications:
[1] A. Dziri, S. Chevobbe, M. Darouich, "Gesture recognition on smart cameras", Proceeding of SPIE 2013, 86590L-86590L-15
[2] C. Bechara, A. Berhault, N. Ventroux, S. Chevobbe, Y. Lhuillier, R. David, D. Etiemble, "A small footprint interleaved multithreaded processor for embedded systems", IEEE International Electronics Circuits and Sydstems (ICECS), 2011, pp. 685-690.

Use of wavelet for image processing in smart cameras with low hardware resources

Research topics : Wavelet, Demosaicing, Denoising, Recognition, Embedded systems

S. Courroux, S. Chevobbe, M. Darouich, M. Paindavoine (LEAD)

ABSTRACT: In this work, we investigate the opportunity to use the wavelet representation to perform good quality image processing algorithms at a low computational complexity. Demosaicing, denoising, contrast correction and classification algorithms are executed over several well-known embedded cores (Leon3, Cortex A9 and DSP C6x). Wavelet-based image reconstruction shows higher image quality and lower computational complexity (3x) than usual spatial reconstruction. The use of wavelet decomposition also permits to increase the recognition rate of faces while decreasing computational complexity by a factor 25.

This work addresses the reconstruction and enhancement of a Colored Filter Array (CFA) image as well as the recognition of badly illuminated faces in the context of low resources cameras. Low and mid-quality processing chains have been designed and executed over different kind embedded processors such as control-oriented, general and DSP-like processors.

Wavelet-based processing chains outperform regular spatial algorithms both in terms of objective quality and computational complexity on low footprint general embedded processors as well as on DSP-like processors. However, in some special cases that have been highlighted, it is possible to process regular spatial-based processing chains faster but at lower quality than wavelet-based processing chains. Consequently, the use of the wavelet representation can help the designer to fit the requirements of the embedded domain.

Eigenfaces face classifier is used to achieve face recognition both in spatial and wavelet domains. In the latter case, only approximation coefficients are classified instead of the whole image. Unevenly face illumination highly degrades recognition rate. Consequently, a contrast correction algorithm is applied prior to the recognition task.

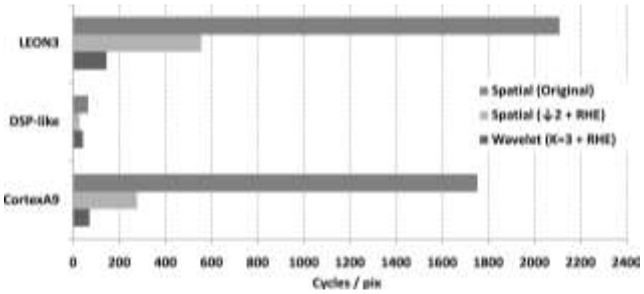


Figure 1: Computational complexity of the different recognition systems, executed over LEON3, Cortex A9 and the DSP-like processor

The computational complexity of three recognition processing chains: naïve, optimal spatial and optimal wavelet, are presented in Figure 1. The classification step requires one multiplication per cycle and per individual in the learning database. Executions on general purpose processors (LEON3 and CortexA9) show that wavelet-based method requires much less cycles than both naïve (15x) and optimal (4x) spatial processing pipelines. The same trends are observed for these two processors since spatial and wavelet processing chains require the same number of multiplication, an operator which is well optimized on the CortexA9. We observe that the length of the input vector has been reduced to 1.56 % of the original image size, while it is 100 % in the naïve case and 25 % in the optimal spatial case.

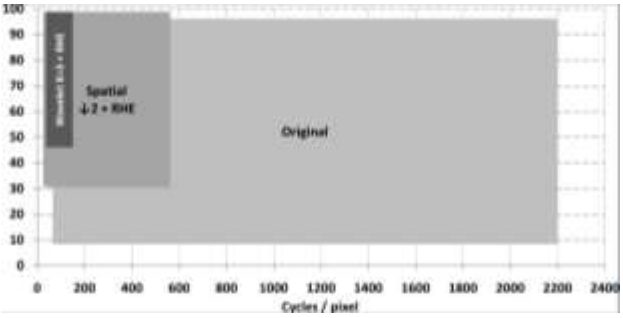


Figure 2: Quality versus computational complexity space of the face recognition application. Wavelet-based processing chain has high recognition rate and requires a low numbers of cycles, whatever the processor is.

Even with a larger vector size, spatial processing chains requires less cycles to execute on the DSP-like processor than wavelet processing chain. Concerning the two spatial chains, the compiler is able to provide a high instruction level parallelism (ILP), which increases performances. The ILP of the classification step is about 7.8 and 2.8 for RHE (Regional Histogram Equalization) algorithm. Consequently, when RHE and classification occupy respectively 9 and 91 % of the total execution time on LEON3, these operations represent respectively 40 and 60 % of the total execution time on CortexA9. As the ILP of the 2D DWT operation is about 1.6, wavelet-based recognition chains does not take advantage of the speedup due to the instruction-level parallelism of the DSP-like processor. The vector size reduction has a non-negligible impact on the memory footprint for the database storage. Indeed, for the optimal spatial case and naïve approach the memory needed is, respectively 600 KBytes and 2.5 MBytes, while it is only 40 KBytes for the wavelet case with 3 levels of decomposition (a reduction by a factor 15 to 64).

Related Publications:
[1] Courroux, S.; Chevobbe, S.; Darouich, M. & Paindavoine, M. (2013), 'Use of wavelet for image processing in smart cameras with low hardware resources ', Journal of Systems Architecture

128 nodes 4.5 mm pitch 15-bit Pressure Sensor Ribbon

Research topics : Aircraft metrology, high resolution pressure measurement, pressure

C.Condemine, J.Willemin, S.Bouquet, S.Robinet, A.Robinet, L.Jouanet, G.Regis*, S.Vitry* (*MIND)

ABSTRACT: Air-flow characterization for aircraft applications, demand high-density and high resolution pressure measurement. A ribbon of 128 15-bits thermally compensated pressure sensors was developed. Each sensor is compounded of off-the-shelf capacitive pressure transducer and an integrated specific sensor interface. The MEMS pressure sensor capacitance variations are converted into digital thanks to a second order switched caps sigma-delta converter. An innovative serial-bit sliding window FIR filter allows asynchronous data transfer. The global system noise is limited to 24 Pa in a range of [20kPa/95kPa] and [-30°C/+60°C].

To optimize a car or aircraft drag and pressure coefficients, both measurement and modeling are necessary. In parallel to mesh optimization in simulation tasks, in-situ advanced metrology must consequently be developed, using pressure monitoring, shape reconstruction, and MEMS sensor integration.

To that purpose, we developed a miniaturized air-flow sensing system to get high pressure measurement density (4.5mm pitch) on an aerodynamic shape [1]. In our specific application case, 128 sensors have to be integrated and interconnected on an existing communication bus. The whole ribbon has stringent application constraints: Firstly, to correlate the pressure data to environmental pressure variations, the pressure offset compensation needs a time resolution better than 1ms. Secondly, to measure the pressure at the same time on the whole ribbon, channel synchronization delay smaller than 400µs is mandatory. Lastly, In order to be compliant with the existing infrastructure, the proposed microsystem must be able to provide a digital output in a frequency range between 200Hz and 1kHz.

An integrated-specific sensor interface was developed to compute a 16-bit, temperature-compensated pressure measurement. The challenges in this work were: compliancy with drastic packaging constraints, high resolution computation data in a low power budget, and innovative data buffering solution, compatible with the fully asynchronous and delay-constrained interrogation protocol.
To optimize the overall power budget and leverage the analog processing, a first 150Hz low-pass filtering stage was realized at mechanical level using the packaging cavity.
In regards to signal bandwidth (BW) and targeted Signal to Noise Ratio (SNR), a switched caps sigma-delta converter is the best solution choice. For a 100Hz bandwidth, a second-order modulator and a 512 oversampling ratio gives a 16 bits resolution. Using a digital channel filtering composed of a Sinc3 (1536 coefficients) and two half band filters (15 and 51 coefficients), the computation latency would have been higher than 150ms with the modulator clock. We thus proposed to use an innovative sliding window Finite Impulse Response (FIR) filter, triggered with the SPI BUS selection (CS) (Fig.1).

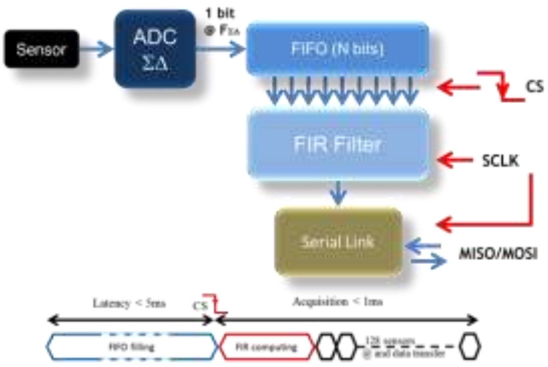


Figure 1 : Pressure sigma-delta architecture

For a 30Pa global precision, we had to compensate the thermal drift mainly due to the sensor (20Pa/°C) The 0.1°C temperature measurement resolution was performed using a 10bits resolution 10 Hz bandwidth sigma-delta converter (OSR=256). A size-optimized filter was designed, composed of a 94 coefficient Sinc3 with a decimator factor of 32, followed by a 16 coefficients Sinc2 decimated by 8.

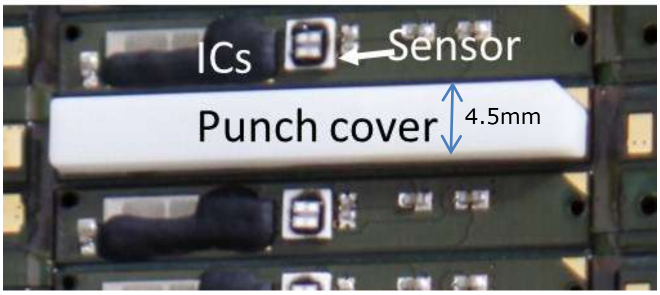


Figure 2 : System on thin ribbon

The circuits were fabricated in 0.35µm CMOS technology. The measured pressure precision is 15 bits and 10 bits for temperature precision, for a sampling frequency between 200Hz and 1kHz and a power consumption of 9.2mW (including bus driving). The global system noise is limited to 24 Pa in a range of [20kPa/95kPa] and [-30°C/+60°C]. The whole ribbon answers at request with less than 1µs maximum delay between channels and 5ms maximum overall latency.

Related Publications :
[1] Condemine, C.; Willemin, J.; Bouquet, S.; Robinet, S.; Robinet, A.; Jouanet, L.; Regis, G.; Compagnon, O. & Vitry, S. (2013), '128 nodes 4.5 mm pitch 15-bit pressure sensor ribbon''Proceedings of the ESSCIRC (ESSCIRC), 2013 Proceedings of the', 229-232.

An Adaptive Output Impedance Gate Drive for Safer and More Efficient Control of Wide Bandgap Devices

Research Topics : Adaptive gate drive, diode-less, HF converter, dead-time loss

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

ABSTRACT: An adaptive gate drive circuit has been fabricated in AMS 0.35µm CMOS process to provide a safer and more efficient control of Wide Bandgap Devices (WBD). The gate driver has an adaptive output impedance for optimal turn-on/off driving conditions and a gate side power transistor switching state detector. Its impedance can be precisely adjusted from 0.7Ω to 12.5Ω during transitions. In a 800 kHz switching frequency diode-less WBD-based synchronous buck converter, the proposed gate drive circuit demonstrates secure but drastic dead-time reduction with a peak performance gain of 20% compared to a fixed dead-time of 50ns.

Wide Bandgap Devices (WBD) like GaN HEMT or SiC JFET offer outstanding performances for high switching frequency, high power density and high temperature applications. WBD can be used into most of converters in place of conventional silicon transistors because they are functionally close; however they are structurally different.

These new devices require more attention on the gate side because of smaller turn-on energy and faster transition times. When current and voltage slew rates become too important destructive overvoltage and faulty turn-on are omnipresent. In such cases by increasing the gate drive impedance, slew rates are limited and the safe operating area is extended. Conversely when transition times are too slow (at low output load) extra switching loss can be reduced by setting the lowest impedance.

Some of WBD, do not have a built-in body diode. In the view of improving synchronous converters' operation a very short dead-time is definitely desired. Secure drastic dead-time reduction is reached by adapting output impedance to balance the impact of the operating point and the temperature on switching characteristics [2] and so maintain the same WBD transition times.

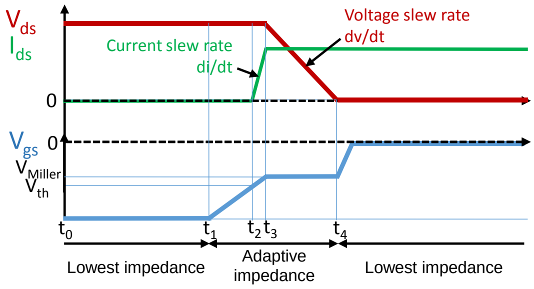


Figure 1 : Turn-on waveforms with adaptive output impedance gate drive

The gate driver adapts its fall and rise output impedance between two switching periods and during steady states by respectively adjusting the number of parallel NMOSs and PMOSs constituting the output buffer. Moreover, a detection of the power device switching state has been implemented into the prebuffer, in order to ensure safest operations. When the gate to source voltage Vgs exceeds the Miller plateau voltage VMiller, the lowest impedance is automatically set as shown on Fig. 1. In such a way the output impedance selection is effective only during turn-on or turn-off and parasitic switching can be avoided.

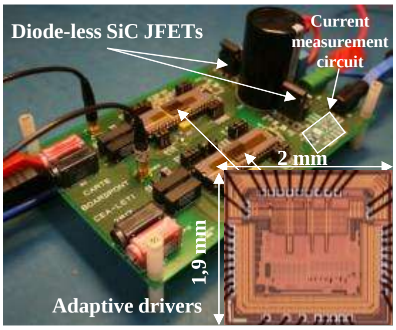


Figure 2 : 800kHz switching frequency diode-less vertical SiC JFET-based synchronous buck converter driven by the adaptive output impedance gate drive circuits

The adaptive output impedance gate drive circuit has been fabricated in AMS 0.35µm HV CMOS (Fig. 2). It can precisely control WBD with adaptive output impedance from 0.7Ω to 12.5 Ω between two switching periods, by adjusting the impedance with an external control unit. Overvoltage due to parasitic inductances into a given synchronous converter has thus been reduced by 80%. Moreover in the diode-less WBD-based synchronous buck converter shown on Fig. 2, the proposed gate drive circuit demonstrates secure but drastic dead-time reduction with a peak performance gain of 20% compared to a fixed dead-time of 50ns.

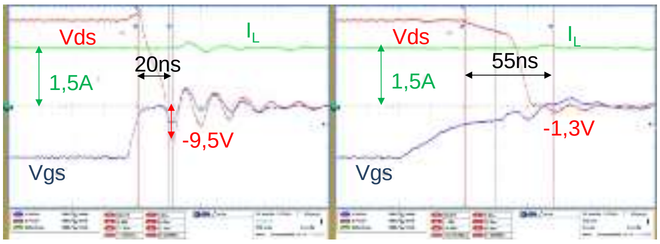


Figure 3 : Turn-on of a SiC JFET driven by the adaptive gate drive circuit into the lowest and the highest output impedance mode: extend safe operating area but extra switching loss.

Related Publications :
[1] R. Grezaud, F. Ayel, N. Rouger, and J.-C. Crebier, "An Adaptive Output Impedance Gate Drive for Safer and More Efficient Control of Wide Bandgap Devices," IEEE Workshop on Wide Bandgap Power Devices and Applications (WiPDA), Oct. 2013
[2] R. Grezaud, F. Ayel, N. Rouger, and J.-C. Crebier, "A Specific Characterization Method for Evaluation of Operating Point and Temperature Impacts on Wide Bandgap Devices," IEEE Workshop on Wide Bandgap Power Devices and Applications (WiPDA), Oct. 2013

A Self-Starting Fully Integrated Auto-Adaptive Converter for Battery-Less Thermal Energy Harvesting

Research Topics : Thermal energy harvesting, adaptive charge pump, battery-less

R. Grezaud, J. Willemin

ABSTRACT: A fully integrated DC/DC converter capable of supplying energy to a battery-less sensor by extracting power from thermoelectric generators (TEGs) over a wide temperature gradient range (3K to 12K) has been fabricated in UMC180nm process. The power management unit of the converter enables to directly power a sensor with a 1.2V regulated voltage output. Under a special operating mode it can also manage larger power consumption operations during a short time. It starts converting at 250mV input voltage and provides up to 1.6mW output power with a 70% peak efficiency using a very low silicon area of 2.86mm².

The power supply of wireless sensors is usually a primary battery which limits the deployment and the life time of the sensor. To reach energy autonomy in environmentally friendly conditions, a battery-free and fully integrated converter optimizing the global power extraction with a variable conversion factor is targeted. Previous work on battery-less thermal energy harvesting report efficient converter, but are not fully integrated.

In [1] we propose a direct path between the scavenger and the application through an efficient converter. Such configuration improves the overall node efficiency [2]. As the wireless sensor is directly powered by the converter, the output voltage has to be regulated to assure stable performance. In the first converter operating mode, the output regulated voltage is set to 1.2V (until t1 on Fig. 1), providing smooth power supply for low power sensing operations. Thr converter can also power the node for a high consuming task (wireless communication) by using a special 1.5V mode. In this second mode, the harvested energy is accumulated in the output capacitor at higher voltage (until t3 on Fig. 1).

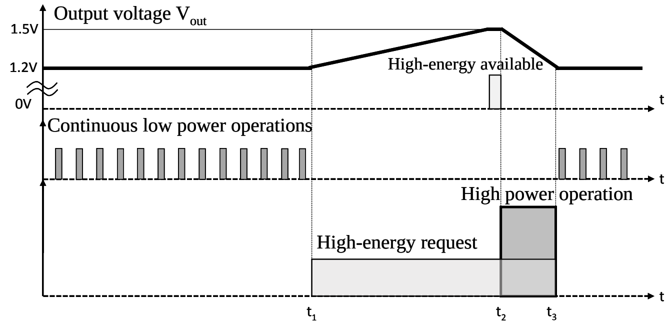


Figure 1 : The Power management unit adapts converter operations according to the available input power and sensor node tasks

The converter was fabricated in UMC180nm process using2.86mm² silicon area (Fig. 2). It is based on a cross-connected Dickson charge pump. By monitoring the available input power, the very low power control circuit adapts the number N of pumping stages to improve the overall efficiency (Fig. 3). The circuit is self-starting when the TEG output voltage is only 250mV. From a 3K to 12K thermal gradient, it provides from 22µW to 1,6mW output power with 70% peak overall efficiency (Fig. 3).

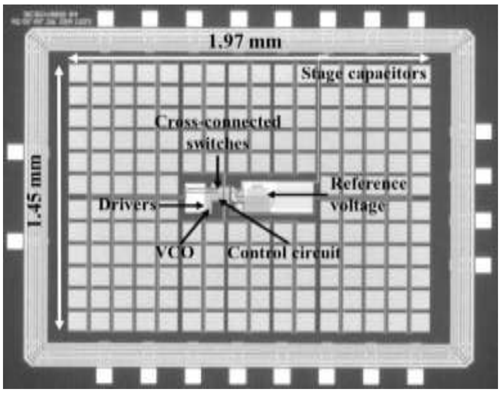


Figure 2 : Fully integrated converter micrograph

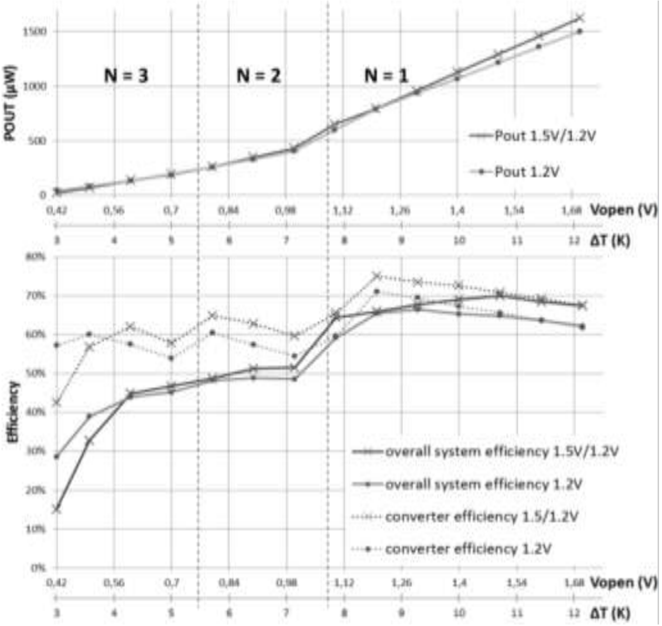


Figure 3: Measured output power, converter efficiency and overall system efficiency. Overall efficiency is the useful output power divided by the maximum recoverable TEG output power

Related Publications :

[1] R. Grezaud, J. Willemin, " A Self-Starting Fully Integrated Auto-Adaptive Converter for Battery-Less Thermal Energy Harvesting," IEEE New Circuits and Systems Conference (NEWCAS), June 2013.

[2] J.-F. Christmann, E. Beigne, C. Condemine, P. Vivet, J. Willemin, N. Leblond, and C. Piguet, “Bringing Robustness and Power Efficiency to Autonomous Energy-Harvesting Microsystems,” IEEE Design Test of Computers, vol. 28, no. 5, pp. 84 –94, Oct. 2011.

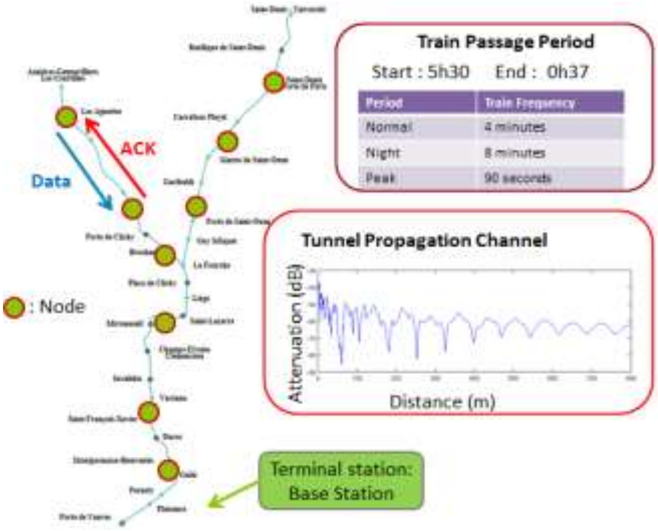
HarvWSNET - A Co-Simulation Platform for Energy Harvesting Wireless Sensor Networks

Research topics : WSN Simulation, Energy Harvesting

A. Didioui, G. Vaumourin, V. Tran, F. Broekaert (Thalès Com), C. Bernier, O. Sentieys (INRIA)

ABSTRACT: These papers [1,2] present the HarvWSNET co-simulation framework developed within the GRECO (GREen Communicating Objects) project. GRECO is a French ANR (National Research Agency) project whose aim is to design an energy efficient wireless platform that is totally autonomous thanks to energy harvesting (EH) capabilities and adaptive power management. The detailed modeling capability of HarvWSNET is illustrated by pre-prototyping a complex wind energy harvesting application for a peer-to-peer subway tunnel wireless sensor network (WSN) application.

The idea of extending WSN node lifetime by equipping each node with an energy harvesting (EH) subsystem has recently attracted a great deal of attention. Indeed, energy harvesting appears to be an ideal candidate for powering WSN nodes: the energy density available in the environment, whether solar, wind, vibrational or thermal in nature, is often compatible with the needs of a WSN application. While energy harvesting appears to be an extremely promising technology for extending WSN lifetime, new tools are required for accelerating time-to-market for these new systems. Indeed, in addition to validating the hardware and software components required by an application, pre-deployment studies must also validate the time-varying availability of the power source in the given application environment.



Period	Train Frequency
Normal	4 minutes
Night	8 minutes
Peak	90 seconds

Figure 1 : Illustration of Subway EH-WSN Application

To this end, the co-simulation platform presented in [1-2] allows for a complete modeling of complex EH-WSN applications thanks to the association of a discrete-events WSN simulator (WSNET) with an industry standard continuous-time simulator (Matlab). This framework is employed to pre-prototype a data-aggregating peer-to-peer WSN consisting of 184 nodes installed at regular intervals of a subway-line tunnel (Fig. 1). Each node is powered by a small wind-turbine which harvests the wind generated by each train passage.

To pre-prototype this application, an accurate model of the wind energy harvester was developed (Fig. 2), including the rectifier, DC/DC converters and supercapacitor. Other models were also defined including models for in-tunnel signal propagation (Fig. 1), train mobility, TDMA-based MAC protocol, data aggregation and acknowledgement, and node power consumption.

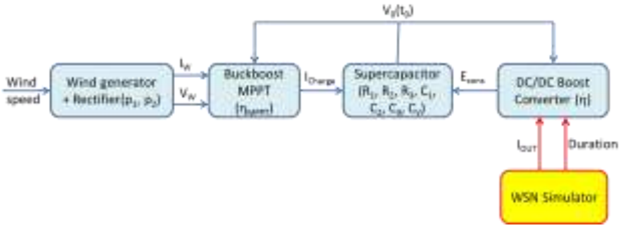

Wind speed
Wind generator + Rectifier(p1, p2)
IW
VW
Buckboost MPPT (ηbypass)
ICharge
Supercapacitor (R1, R2, R3, C1, C2, C3, C4)
Econs
DC/DC Boost Converter (η)
VS(t0)
IOUT
Duration
WSN Simulator

Figure 2 : Wind Energy Harvester Model

Simulation results show the time-varying behavior of each node's supercapacitor voltage (Fig. 3) which tends to recharge when train passage frequency is high and tends to discharge during the last 5 hours of the simulation when trains are at a stop. These simulations are used to evaluate the viability of this application, identify appropriate network architecture and determine sensor activity period, data delay and network autonomy in case of train traffic interruption. In addition, the importance of network and node-level power-aware protocols is illustrated since the network must adapt to the time-varying nature of the power source. Finally, the simulations reach a quick verdict as to the viability of this application, hence limiting the prototyping costs and delay and minimizing the time-to-market of novel EH-WSN scenarios.

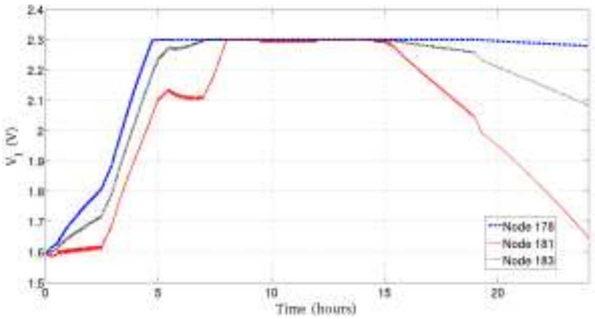


Figure 3 : Supercapacitor voltage during a 24 hour simulation

Related Publications
[1] A. Didioui, C. Bernier, D. Morche, O. Sentieys, "HarvWSNet: A co-simulation framework for energy harvesting wireless sensor networks", 2013 International Computing, Networking and Communications Conference (ICNC).
[2] F. Broekaert, A. Didioui, C. Bernier, O. Sentieys, "Prototyping an Energy Harvesting Wireless Sensor Network Application Using HarvWSNet", ARCS 2013.

Ultra Wide Band : Ultra-Fine and Robust Localization with Quadrature Receivers

Research topics : UWB, Localization, Beamforming, Antennas

D. Morche, G. Masson, S. De Rivaz, F. Dehmas
S. Paquelet, A. Bisiaux (Mitsubishi), Fourquin, Gaubert, Bourdel (IM2NP)

ABSTRACT: In this project, we have demonstrated that high localization precision can be obtained using quadrature receivers. The down-conversion to baseband is exploited to filter out any unwanted signal. By combining RF filter, 5th order gm-CT filter, window integration and coherent integration, rejection higher than 80 dB can be achieved. Then the correlation with an orthogonal basis is exploited to reach very good ranging precision (few cms). Phase information can be used to improve even more the precision (few mms)

Localization technologies are facing a growing interest in the industry since they can offer a wide variety of applications. Among the different technologies, UWB is known to potentially offer the best precision. Whereas However, up the now, the performances of the existing solution were limited in terms of performances.

At the very beginning of UWB, most of the IR-UWB localization solutions were based on non-coherent receiver with poor performances. Since almost ten years, the interest in coherent architectures has been growing in order to reach better performances with improved robustness to out-of band blocker which is one of the main issues in Ultra Wide Band due to spectrum saturation.

In this project [1], we have targeted low power, long range and high robustness localization applications. For that purpose we have selected a double quadrature receiver architecture which is presented in Figure 1 (without the frequency synthesis). This architecture enable us to use a low rate sampling frequency while keeping very good precision in the time of arrival precision. The second oscillator frequency is chosen to be a sub-multiple of the first one to simplify the PLL design.

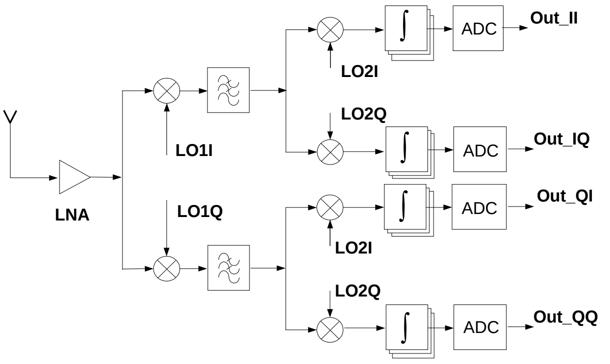


Figure 1 : LORELEI Receiver Architecture

To reach high precision performances and long range, a 1.5 GHz bandwidth has been adopted. However, the receiver architecture is compatible with classical 500 MHz channel bandwidth. It can even switch from one channel to another without modifying the PLL frequency.

Then a classical gmC filter has been design to attenuate out of band signals by 20 dB using the well-known Nauta Cell. Following this filter, the signal is integrated during a dedicated window with a switched capacitor infrastructure. Lastly, the signal is converted from analog to the digital domain with a low power flash ADC. The whole circuit has been design using 130nm CMOS technology from STMicroelectronics.

New metrics have been developed to extract the time of arrival of the received pulse. The ranging error is now lower than 1.5 cm as presented in Figure 2. Much better precision can be obtained by exploiting the phase information of the received signal. In that case, error is lower than 2,5mm but the precision becomes dependent of the antenna characteristics.

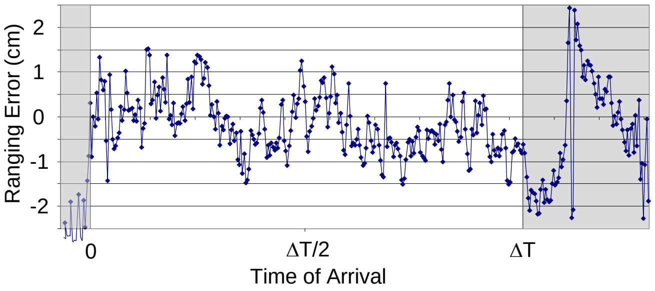


Figure 2 : Ranging error of the Time of Arrival Estimation as a function of the arrival time in the window

The obtained sensitivity is better than -95 dBm@1Mb/s justifying the architecture choice [2]. The 50mW power consumption is fully compatible with mobile devices. The proposed solution provides the lowest sensitivity and the best ranging precision. Since the architecture is fully coherent, the range (300m) can be extended even more by resorting to multi-antenna architectures. The performances of the localisation can be even more improved by extracting the angle of arrival of each path which has been shown to be possible with the proposed architecture [3].

This development paves the way for the industrialization of UWB localisation solutions.

Related Publications:
[1] D.Morche, G.Masson, S.De Rivaz, F. Dehmas S.Paquelet, A.Bisiaux, O.Fourquin, J.Gaubert, S.Bourdel "A Double Quadrature UWB Receiver for wide Range Localization Applications with sub-cm Ranging Precision" IEEE Journal of Solid-State Circuits 2013, Volume 49, Issue 10, October 2013
[2] D.Morche, M.Pelissier, G.Masson, P.Vincent "UWB : Innovative Architectures Enable Disruptive Low Power Wireless Applications » Invited paper in DATE Conference 2012
[3] Farid Bautista, Dominique Morche, Serge Bories, Gilles .Masson "Antenna Characteristics and Ranging Robustness with Double Quadrature Receiver and UWB Impulse Radio" ICUWB 2012

A low power 60-GHz 2.2-Gbps UWB transceiver with integrated antennas for short range communications

Research topics : mmW, UWB, CMOS, integrated antenna, wireless

A. Siligaris, F. Chaix, M. Pelissier, V. Puyal, J. Zevallos, L. Dussopt, and P. Vincent

ABSTRACT: A 60-GHz low power fully integrated transceiver including antennas, fabricated in CMOS 65nm SOI and packaged in low cost QFN is presented. The circuit achieves 2 Gbps and 500 Mbps rates at 7.5 cm and 22.5 cm transmission ranges respectively. The transceiver energy efficiency is lower than 50 pJ/bit thanks to scalable power consumption using pulse generator and Super Regenerator Oscillator architecture.

Future consumer mobile platforms need high rate wireless connections for data exchange and video streaming. 60-GHz band has been intensively explored for such purpose, but existing solutions exhibit high power consumption. This work describes a fully integrated 60 GHz low power transceiver, including the antennas, fabricated in a CMOS 65nm SOI technology. The chip is packaged in a low cost QFN pre-molded cavity. The circuit targets mobile devices requiring multi-Gbps wireless data transfer for pear-to-pear exchange, data synchronization and kiosk application. It can also be used in chip-to-chip or board-to-board wireless communications where low power consumption is required. Thanks to low complexity in the transceiver architecture and by using On-Off-Keying (OOK) modulation scheme, lower than 100 mW is attained for the whole communication at 2 Gbps throughput.

The transmitter is based in a switched pulse-injected oscillator. The input digital data to be transmitted are injected serially in the oscillator and a pulsed sine signal with center frequency at 61 GHz is generated. The pulse is next amplified by a single stage power amplifier and radiated by the integrated antenna. At the receiver side, the system uses a Super Regenerator Oscillator (SRO). It offers many advantages for wideband pulsed oscillating signals in terms of duty cycling capability, sensitivity bandwidth, and instantaneous gain. Synchronization of the SRO is performed with integrated DLLs. Finally an envelope detector and a comparator are used to decide if a "1" or a "0" is detected. Integrated antenna design takes into account the global chip environment like wire bonds, circuit and packaging metallic elements as well as the polymer material properties.

Wireless links are tested with two boards containing fully integrated transceivers placed in an office like environment (figure 2). For a BER ≤ 10-5, a 7.5 cm distance is achieved for 2 Gbps rate and up to 22.5 cm distance for 500 Mbps. This work shows an excellent trade-off between wireless transmission range, data rate and power consumption achieving better than 50 pJ/bit total energy efficiency, which is the best to our knowledge and sets the state-of-the art for this type of transceivers. The proposed circuit provides a fully integrated and low-cost solution for high data rate, low power, and short range wireless data exchange chipsets.

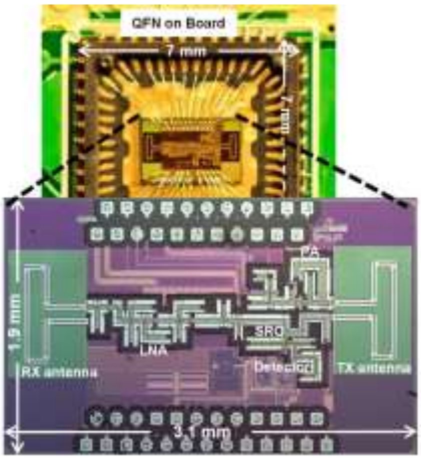


Figure 1 : Chip-in-package photograph (open lid) and details of the chip. Dimensions: 3.1x1.9 mm².

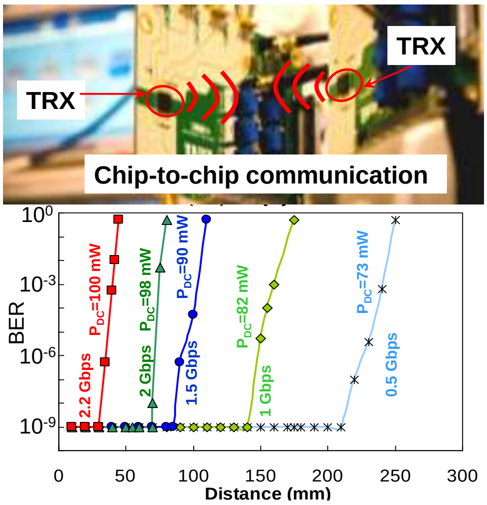


Figure 2 : Experimental setup for wireless link and BER versus communication range for various data rates indicating the total transceiver dissipated power for each case

Related Publications:
[1] A. Siligaris, F. Chaix, M. Pelissier, V. Puyal, J. Zevallos, L. Dussopt, and P. Vincent, "A low power 60-GHz 2.2-Gbps UWB transceiver with integrated antennas for short range communications," IEEE Radio Frequency Integrated Circuits Symposium, pp. 297-300, June 2013.
[2] J. Zevallos Luna, L. Dussopt, and A. Siligaris, "Hybrid On-Chip/In-Package Integrated Antennas for Millimeter-Wave Short-Range Communications," Transactions on IEEE Antennas and Propagation, vol. 61, no. 11, pp. 5377-5384, Nov. 2013.
[3] L. Dussopt, J. Zevallos Luna, and A. Siligaris, "On-chip/in-package integrated antenna for millimeter-wave medium and long-range applications'' International Workshop on Antenna Technology, (iWAT), pp. 203-206, March 2013.

Inductor shielding strategies to protect mmW LC-VCOs from high-frequency substrate noise

Research topics : CMOS RF integrated circuits, temperature sensor, testing, self-healing

J.L. González, M. Molina (UPC), X. Aragonés (UPC) D. Mateo (UPC)

ABSTRACT: High frequency signals coupling through the substrate is the main source of VCO pulling in fully integrated transceivers [1]. In this work we investigate and compare two shielding strategies commonly used to avoid this effect: grounded shields and floating shields. The results of our investigations show that, even if the coupling from the substrate is very similar in the two cases, the final effect on the VCO is strongly reduced by the floating shield inductor due to its larger quality factor which reduces the ability of the coupled interferences to pull the oscillator.

The use of grounded shields (GS) under integrated inductors is a common practice today. It allows to provide a controlled environment to the inductor and to reduce the coupling of undesired signals from the substrate. We show, however, that using a floating shield (FS) is an even better idea regarding the resilience of the VCO to interference signals coming from the substrate [2]. Both types of inductors are illustrated in Fig. 1. The inductors where sized for a 60 GHz VCO [3] shown in Fig. 2. The prototype IC includes a pad contacting to the substrate through a large line of p+ contacts all along the right edge of the circuit (see Fig. 2). This pad is used to inject an interference signal into the substrate at a frequency close to the oscillator frequency. The power of the signal and its offset with respect to the VCO oscillator is set so that no pulling is observed. This allows measuring the spur appearing at the output of the VCO using a signal analyzer.

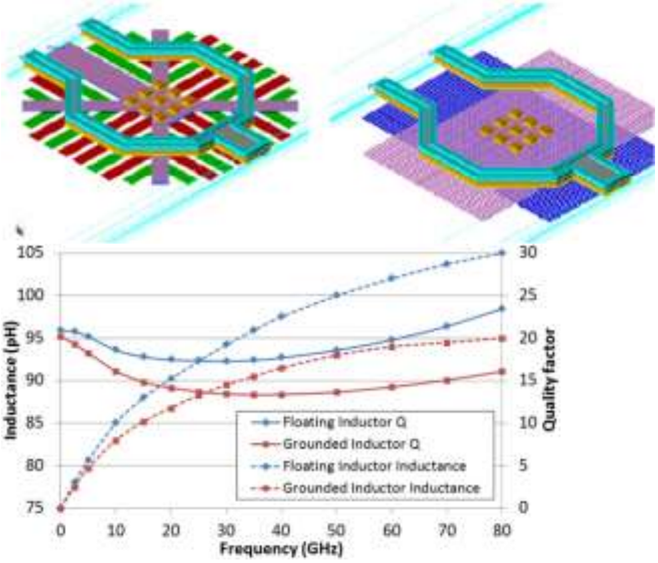


Figure 1: Integrated inductor with grounded (left) and floating (right) shield, and comparison of quality factors and inductance value.

First of all, we have done 3D EM circuit simulations and we have found that the coupling from the substrate to both inductors is very similar.

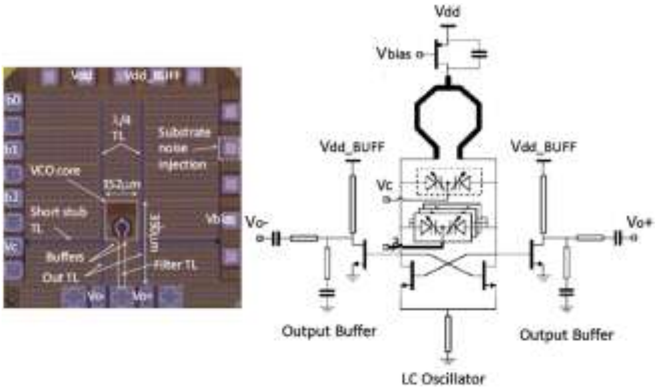


Figure 2: 60 Ghz VCO using either type of inductor.

Secondly, an experimental analysis of substrate interference onto the VCO has been done [2]. The measurements results are shown in Fig. 3. Two extreme values of control voltage Vc where used, which correspond to the extreme values of quality factor for the tank using either of the inductors (GS:7.1-9.7, FS:8-11). The "grounded" shield inductor shows large spur amplitudes compared to the floating shield inductor for the high Vc value, which corresponds to the case where the overall tank quality factor is dominated by the inductor. For lower Vc values, the varactor quality factor, which is the same in both cases, plays a more important role. Therefore we can conclude that the higher quality factor of the FS inductor VCO is the reason of its better resilience to substrate coupled interference.

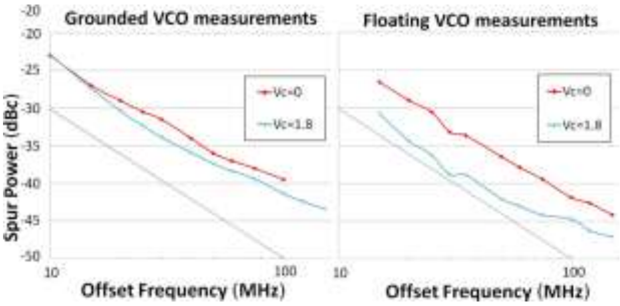


Figure 3: Relative amplitude of the output spur for a 10dBm input tone at the substrate as a function between the VCO and the substrate signal frequency offset.

Related Publications:
[1] M. Molina, D. Gómez, X. Aragonés, D. Mateo, J.L. González, "Design of a 2.5 GHz QVCO robust against high frequency substrate noise," Microwave and Optical Technology Letters, Vol. 53, No. 7, pp. 1632–1637, July 2011.
[2] Marc Molina, Xavier Aragonés, Diego Mateo, José Luis González, "Inductor shielding strategies to protect mmW LC-VCOs from high frequency substrate noise," Microelectronics Journal, Vol. 44, No. 5, pp. 405-413, May 2013.
[3] J.L. Gonzalez, F. Badets, B. Martineau, D. Belot, "A 56-GHz LC-Tank VCO With 17% tuning range in 65-nm bulk CMOS for wireless HDMI, " IEEE Transactions on Microwave Theory and Techniques, Vol. 58, No. 5, pp. 1359-1366, May 2010.

A 285 GHz sub-harmonic injection locked oscillator in 65nm CMOS

Research topics : mmW, sub-THz, VCO, CMOS, locking

J.M. Guerra, A. Siligaris, J.-F. Lampin (IEMN), F. Danneville (IEMN), P. Vincent

ABSTRACT: 285 GHz Sub Harmonic Injection Locked Oscillator (SHILO) is presented using a standard 65nm CMOS process. The architecture of this oscillator is based on the differential LC tank with push-push but adapted to obtain a third harmonic oscillation. The output power is -19 dBm at 285 GHz for a dc power of 70 mW. This oscillator offers a measured phase noise of -96.3 dBc/Hz at 10 MHz and -80.5 dBc/Hz at 1 MHz, and a tuning range from 284.2 GHz to 289 GHz. The SHILO can be locked all along the tuning range with an injection signal corresponding to one sixth of the output frequency. The chip size is 921x451 µm².

Sub-THz frequencies are explored for applications such as THz imaging and high throughput rate short-range wireless communications. This work aims to explore the feasibility of one of the most challenging blocks in such systems: the local oscillator. Indeed, designing oscillators operating close or at higher frequencies than the cut-off frequencies of the transistors needs new designing techniques. The common technique is sub-harmonic generation, which means that a fundamental tone is synthesized below the cut off frequency and higher harmonics are boosted using push-push, triple-push or quadruple-push techniques. Here, a modified version of the classic scheme of second harmonic generation is presented in order to exploit a third harmonic. Moreover, the designed and fabricated oscillator aims at its integration into complete frequency synthesizers. For that, a sub-harmonic injection signal is used in order to lock the oscillator and thus to generate a 285 GHz output signal that copies the reference signal frequency and phase noise.

The architecture of the sub harmonic injection locked oscillator is presented in Figure 1. A differential oscillator generates the fundamental signal (part A); the second harmonic is obtained using a quarter wavelength transmission line (push-push technique). The injection transistors (part B) are used for locking the oscillator in order to stabilize the oscillation frequency and fix the phase noise on the injection source. The injection signal (around 47 GHz) is one sixth of the output oscillation frequency. The third harmonic is generated by a single balanced differential active mixer (part C) that mixes the fundamental tone and the second harmonic obtained by push-push (push-push & mix). Output 50 Ohm matching is adjusted to 285 GHz using microstrip transmission lines and an output balun for single ended measurements. Thus, lower frequencymixing products are rejected.
Measurements were carried out under probes. Figure 2 shows the free running oscillation frequency that spans from 284.2 GHz to 289 GHz while the output power varies from -19 dBm to -27 dBm. We observe that simulations fit measurements with good accuracy which validates the design methods and models.

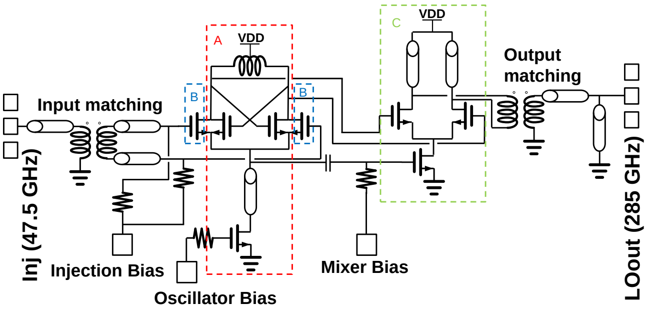


Figure 1: 285-GHz oscillator architecture.

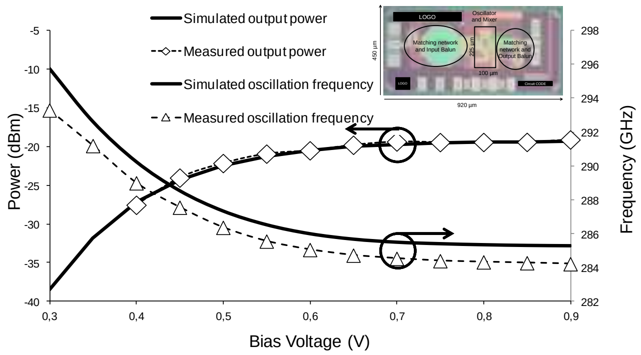


Figure 2: Output power and free running oscillation frequency vs. bias voltage. Symbols: measurements. Line: simulation. Inset: chip micrograph.

This work shows that it is possible to synthesize signals well above cut-off frequencies using specific design techniques. It is the first oscillator that offers the possibility to create a sub-THz signal that is locked on a lower frequency reference. It paves the way for integration in a complete transceiver communication system.

Related Publications:
[1] J.M. Guerra, A. Siligaris, J.-F. Lampin, F. Danneville, P. Vincent, "A 285 GHz sub-harmonic injection locked oscillator in 65nm CMOS technology," IEEE MTT-S International Microwave Symposium Digest (IMS), pp. 1-3, June 2013

A 283 GHz low power heterodyne receiver with on-chip local oscillator in 65 nm CMOS process

Research topics : mmW, sub-THz, heterodyne receiver, CMOS, locking

J.M. Guerra, A. Siligaris, J.-F. Lampin (IEMN), F. Danneville (IEMN), P. Vincent

ABSTRACT: A Fully integrated 283 GHz heterodyne receiver in 65 nm CMOS process is presented in this paper. The circuit includes a resistive differential mixer, an intermediate frequency amplifier and a 282 GHz sub-harmonic injection locked oscillator. The on-chip oscillator generates a 94 GHz fundamental tone but exploits a 282 GHz third harmonic. An injection signal of 47 GHz (one sixth of the RF frequency) is used to lock the oscillator on a reference. The receiver measured conversion gain is -6 dB for a DC power consumption of 97.6 mW. Simulated noise figure is 38 dB.

Sub-THz frequencies are explored for different applications such as mmW imaging for security or biomedical purposes, high data transfer and compact range radar. This work aims to explore the feasibility of a coherent heterodyne receiver that operates well above ft and fmax of the transistors. This challenging objective is achieved through appropriate design techniques that allow fabricating a coherent heterodyne receiver operating at 283 GHz in a standard CMOS 65nm process.

The receiver is constructed by exploiting the advantages of MOS transistors in various regimes of operation and frequency bands. Thus, knowing that it is not possible to obtain positive gain at the 280-GHz band, the RF signal is directly down-converted to low frequency for amplification. At this point, the down-conversion mixer uses a passive structure in which a frequency conversion is possible without high conversion losses even at very high frequencies. It uses a typical characteristic of FET transistors, that is, the resistive mixing technique. Indeed, this technique is not feasible in bipolar technology. The down-converted signal is then simply amplified with a low-noise, large-band base-band amplifier. Thus, the main challenge is carried to the local oscillator signal (LO) that is pumping the passive mixer. Figure 1 illustrates the 283-GHz receiver architecture.

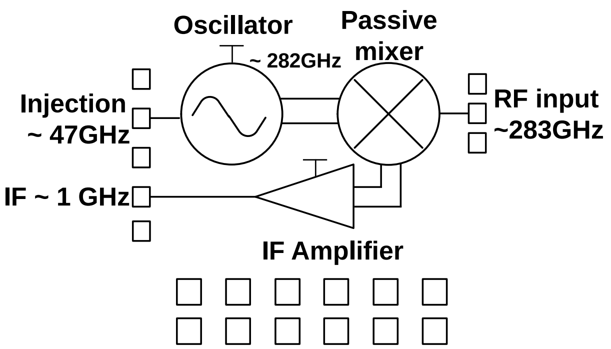


Figure 1 : 283-GHz receiver architecture.

In order to render the receiver coherent (amplitude and phase detection of the signal), the oscillator must provide a stable frequency with low phase noise. In other words, the oscillator must be locked on a low-frequency reference signal. This is achieved with the use of a sub-harmonic locking technique around 47 GHz. Hence, the LO output signal at 282 GHz is a multiple of a 47-GHz reference that copies frequency and phase characteristics. The oscillator core fundamental oscillation frequency is designed to be around 94 GHz. It uses a triple-push architecture that allows combining and extracting the third harmonic, i.e. the 282 GHz LO signal [2].

Measurements were carried out under probes. The circuit was locked with an external frequency synthesizer around 47 GHz. The conversion gain is -6 dB with in both upper and lower side bands and the bandwidth is more than 1-GHz. Simulated noise figure is around 38 dB. The total power consumption of the circuit is 97 mW under 1.2 V supply.

This work shows the feasibility of CMOS receivers that operate well above transistors cut-off frequencies. Appropriate receiver architecture choice and design techniques using harmonic boosting show that heterodyne coherent transceivers can be made in standard CMOS at sub-THz frequency band.

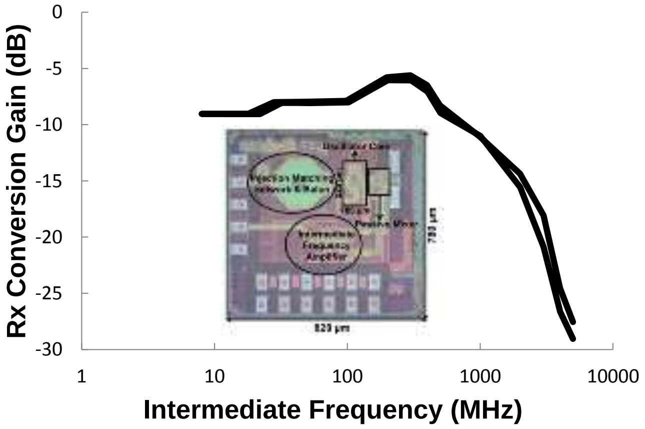


Figure 2 : Receiver conversion gain vs. IF frequency for a fixed LO frequency of 282.44 GHz (lower and upper side bands). Inset: chip micrograph.

Related Publications:
[1] J.M. Guerra, A. Siligaris, J.-F. Lampin, F. Danneville, P. Vincent, "A 283 GHz low power heterodyne receiver with on-chip local oscillator in 65 nm CMOS process," IEEE Radio Frequency Integrated Circuits Symposium (RFIC), pp.301-304, June 2013
[2] J.M. Guerra, A. Siligaris, J.-F. Lampin, F. Danneville, P. Vincent, "A 285 GHz sub-harmonic injection locked oscillator in 65nm CMOS technology," IEEE MTT-S International Microwave Symposium Digest (IMS), pp. 1-3, June 2013

Approximated solution of Van der Pol equation describing transition from forced to free mode of oscillation

Research topics: Oscillators, Injection locking, Van Der Pol, UWB

C. Jany, A. Siligaris, M. Zarudniev, P. Ferrari (IMEP-LAHC)

Abstract: A compact expression describing the transient behavior of the Van der Pol oscillator is presented. An original simplification on the amplitude variation leads to a new expression of the oscillation phase transient behavior. This new expression is particularly suitable for describing oscillator relaxation, i.e. when the oscillator switches from a forced to a free oscillation state. It is shown that by choosing the injection stop instant correctly, the free running frequency can be achieved instantaneously.

The analytical study of oscillators has raised a lot of interest in the past century, with the aim of understanding the behavior of the first electrical oscillators built in the early 1900s. The conventional theory provides a solution for both the phase and the amplitude of weakly forced oscillators (an example of such an oscillator is shown in Fig.1b), but only for the amplitude of free running oscillators (see Fig.1a). In this work an extended analysis based on Krylov-Bogoliubov's method of averaging is proposed from which a solution for both the amplitude and the phase of free oscillators can be found. The novel analysis enables the study of the transition from the forced to the free oscillation state for the Van der Pol oscillator. This work is part of a broader theoretical study of forced oscillators driven by pulsed sinusoidal signals.

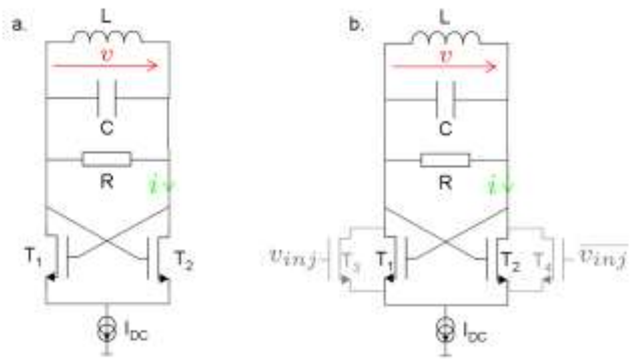


Figure 1: Oscillator composed of a cross-coupled nMOS transistor pair and a LC tank (Fig.1a). Same oscillator with it's injection network (Fig.1b).

Assuming that the transistor pair current-voltage characteristic is of Van Der Pol type, the circuit of Fig.1a can be described by :

$$\ddot{v}+\frac{1}{LC}v=\frac{1}{C}\left(\alpha-\frac{1}{R}-\gamma v^{2}\right)\dot{v}$$

(1)

This is the free Van Der Pol equation, which finds no exact analytical solution in the literacy. The Krylov-Bogoliubov's method of averaging provides a solution for the amplitude of oscillations, but no solutions for the phase. In this work, a model of the amplitude solution leads to a phase approximated solution (2) and thus to an approximated solution for the Van Der Pol equation.

$$\varphi(t)=\arcsin\left(\frac{\sin(\varphi_0)}{1+\dfrac{A_F}{A_0-A_F}}\left(\frac{A_F}{A_0-A_F}+e^{-\varepsilon\omega_0 t}\right)\right) \qquad (2)$$

This solution is validated in the case of small variation between initial and final amplitudes (Fig.2), which suit for the understanding of the transition between the forced and the free oscillation in the case of weakly forced oscillator.

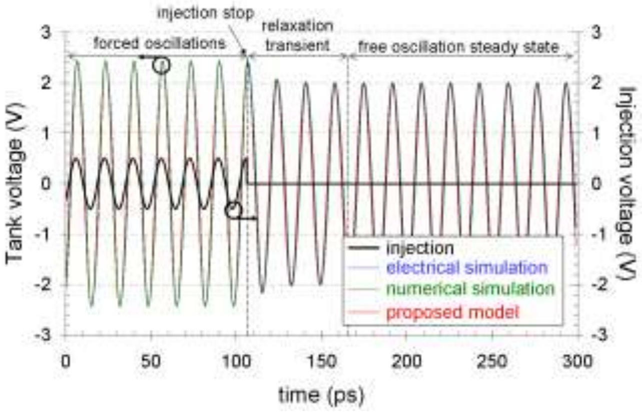


Figure 2: Transient behavior of the oscillator when injection stops at arbitrary moment.

This analysis allows a deeper understanding of the operation of locked oscillators when the injection signal is non-continuous, for example in the case of super-regenerative receivers for UWB impulse radio signals. The equations presented in this letter can be used in this type of systems for sizing the injected oscillator in order to fulfill some transient requirements.

Related Publications :
[1] C. Jany, A. Siligaris, M. Zarudniev, ”Approximated solution of Van der Pol equation describing transition from forced to free mode of oscillation,” IET Electronics Letters, vol. 49, no. 13, pp. 786-787, June 2013.

Network internal signal feedback and injection: Interconnection matrix redesign

Research topics: PLL interconnection, automatic control, convex optimization

M. Zarudniev, P. Villard, G. Scorletti (AMPERE), A. Korniienko (AMPERE)

ABSTRACT: The design of network (i.e. interconnection) of identical subsystems emerges in various engineering fields, with some open issues. One of them is how to "retune" the interconnection in order to ensure the stability and the performance of the global system. Based on the LFT representation and on the input-output framework, we propose some efficient "retuning" methods using convex optimization involving LMI constraints. The proposed approach can be interpreted as an extension of usual state-space methods. Its application is examplified on the design of a network of PLLs.

In Automatic Control, a popular and successful paradigm for Linear Time Invariant (LTI) systems is the state space representation approach. In this approach, a large number of efficient analysis and synthesis methods were obtained using matrix computation and more recently convex optimization over Linear Matrix Inequality (LMI) constraints. Another interest is the physical realization of a state space model as a block diagram involving integrators and constant gains. For the design of systems, the well-known state feedback control paradigm leads to an interesting interpretation: for such a system, how to retune some gains in order to achieve stability and fulfill technical requirements (performance).

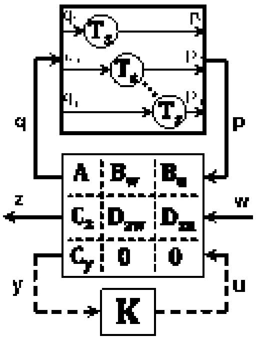


Figure 1: LFT graphical modeling

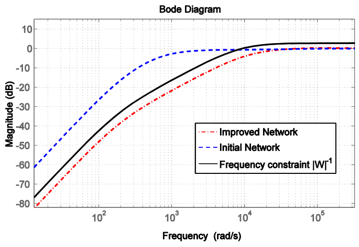


Figure 2: Performance improvement

Recently, a strong interest emerged in Microelectronics for the design of networks of Phase Locked Loops (PLL), where the PLLs are identical [1]. The purpose is to achieve the synchronization of the PLLs with the design specifications formulated in terms of frequency constraints (see Figure 2). In this context, the use of the H-infinity norm is convenient. Note that this problem more generally pertains to the synchronization of oscillators [2]. These networks can be interpreted as block diagrams involving constant gains K to be tuned and identical dynamical LTI systems (see Ts blocks on Figure 1). These dynamical systems are usually different from integrators.

We propose an extension of some feedback synthesis methods usual for the LTI state-space approach, to the case of models which can be realized as block diagrams involving a matrix K of constant gains, in the sequel referred as to the interconnection and dynamical LTI systems, referred as to the subsystems. The proposed methods are efficient, since they are based on convex optimization. With the proposed framework, we prove that it is possible to use convex optimization in order to address some control problems which are not convex when formulated in the state-space representation formulation [3]. Thus, the direct design of the interconnection of identical systems based on the frequency constraints from technical specification becomes possible (see Figure 2).

To this purpose, we use the Linear Fractional Representation modeling usually named LFT modeling. This modeling technique allows representing general block diagrams, including the block diagrams corresponding to state-space representations. Though the general framework has been largely investigated from the 90's, its potential interest is still largely unexplored, even if many interesting results were obtained. One of our contributions [3] is the application of this framework to the design of systems expressed as the interconnection of subsystems. In our previous work, we focused on the design of the subsystems in order to ensure a performance from technical requirements for the (overall) system. We now give the systematic "retuning" of the interconnection in order to improve the system performance. A similar problem was considered in the state-of-the-art approaches with a strong emphasis on the performance analysis. Nevertheless, in contrast with our synthesis approach, the authors gave only some recommendations for the interconnection retuning.

Related Publications:
[1] A. Korniienko, E. Colinet, G. Scorletti, E. Blanco, D. Galayko, and J. Juillard, 'A clock network of distributed adplls using an asymmetric comparison strategy', 2010 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 3212–3215, May-Jun. 2010.
[2] M. Zarudniev, E. Colinet, P. Villard, U. Ebels,M. Quinsat, and G. Scorletti. 'Synchronization of a spintorque oscillator array by a radiofrequency current'. Mechatronics, 22(5) :552 - 555, 2012.
[3] M.Zarudniev, A.Korniienko, G.Scorletti, P.Villard, 'Network Internal Signal Feedback and Injection : Interconnection Matrix Design', 52st IEEE Conference on Decision and Control 2013

Yield Optimization
at System Level and Circuit Level with Redundancy

Research topics : Variability, System, Yield, Optimization, Redundancy

D. Morche, A. Oguz, F. Enikeeva (LJK), S. Nazin (LJK)

ABSTRACT: In this study, we have shown that the impact of the variability can be reduced by considering its impact at the system level rather than at the building blocks level. The performances of the receiver have been optimized thanks to a stochastic gradient iterative procedure. Thanks to this procedure, the yield has moved from 35 % to 82 %. Afterwards, we have shown that thanks to redundancy it is possible to reduce simultaneously power consumption and area while improving the performances.

In these deep sub-micron technologies, variability is becoming more and more important. It is becoming one major issue for technology scaling. New approaches at CAD level are required to circumvent this problem.

Nowadays, for complex systems such as RF receivers, the yield constraint is usually applied at building blocks level. The overall constraints are shared between these blocks and each designer is doing his best to reach the targeted yield. However, since the individual performances of the blocks are highly correlated, high benefit can be taken from a global optimization.

The approach has been applied to a receiver developed in BiCMOS technology for 24 GHz radar applications. It is composed of 8 blocks : 3 LNA stages, Mixer, 2 Filter stages and an amplifier followed by some external stages as depicted on Figure 1.

To achieve the optimization, we resort to meta-modeling of all building blocks in order to link the technology characteristics and the performances of the considered blocks. Instead of developing our own modeling technique, we have employed DoE approach, which is more and more used in the industry to replace the time consuming Monte-Carlo simulations.

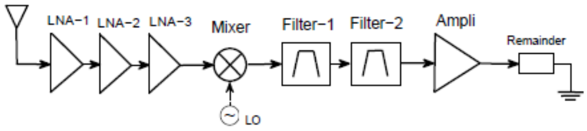


Figure 1 : Receiver Architecture from [1]

Each building block has got tuning parameters which permit to slightly modify its characteristics. The performances of the whole chain are derived from the well-known Friis relationships but they could also have been obtained through simulations. Then, the overall yield has been optimized thanks to stochastic gradient iterative. After 300 iterations, the initial yield of 35% is increased up to 85% which is more than twice higher [2].

Some redundancy improvement method has been also developed at the building block level by exploiting redundancy. The basic idea is to duplicate elementary devices (transistor, capacitors,…) and to select the more matched ones. As matching is particularly significant in ADC performances, the proposed approach has been applied to a SAR ADC. The benefit in terms of matching precision is so high that it permits to drastically reduce the size of the capacitors. Figure 2 shows the influence of the number of redundant capacitors used for the elementary capacitor on the yield of the ADC. These curves have been obtained from statistical simulations. As we target an rms error due to capacitor mismatch equal to 10% of the quantization step, selecting 2 capacitors out of 4 is enough to improve the yield from 40% up to more than 99,9%.

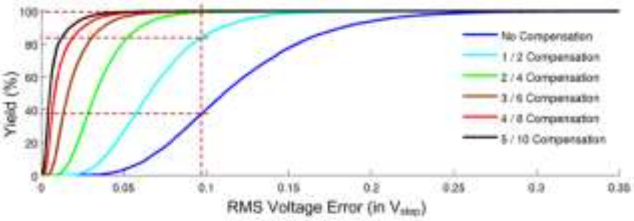


Figure 2 : Influence of the redundancy scheme on the ADC yield as a function of the targeted RMS Voltage Error

This shows that with such a scheme, the size of the elementary capacitor can be modified thus modifying the relationship between the yield and the redundancy scheme. However, extracting new curves is a complex procedure since statistical simulations are needed. With such an approach, the optimization is a complex and tedious process.

To solve this challenge, we have derived an analytical expression [3] linking the yield and the elementary area. Then, by using the normal approximation of the yield, we have proposed a faster way of solving the total area optimization problem. This approach can help to select an optimal number of redundant components. We have also shown that the redundancy method can be applied to simultaneous reduction of the cost (total area) and the power consumption (component size) of advanced systems. With this approach, the routing area can also be taken into account for the global optimization.

Related Publications:
[1] L. Moquillon et al., “Low-Cost Fully Integrated BiCMOS Transceiver for Pulsed 24-GHz Automotive Radar Sensors”, in CICC-08, pp. 475-478
[2] Sergey Nazim, Dominique Morche, Alexandre Reihnhardt "Yield Optimization for Radio Frequency Receiver at System Level" DATE' 2012
[3] F.Enikeeva, D.Morche, A.Oguz “'Yield Improvement by the Redundancy Method for Component Calibration” Electronic Letters, Volume 49, Issue 13, pp. 1-2, 20th June 2013

Process variation compensation for PLL in FDSOI 28nm

Research topics : FDSOI, PLL, Phase Noise, Spurious Calibration

A. Fonseca, E. De Foucauld, P. Lorenzini (EPIB-UNS), G. Jacquemod (EPIB-UNS)

ABSTRACT: This paper presents process variation compensations applied to Ultra Low Power PLL, in order to decrease spurious level created by switching phase divider architecture. The Voltage Controlled Ring Oscillator is designed in FDSOI 28nm and optimized for Bluetooth application.

The local oscillator (LO) in radiofrequency (RF) wireless communications is critical to get low bit error rate (BER) resulting in good transmission quality. Therefore the communication standard - in our case Bluetooth (BT) - sets transmitter specifications, from which result frequency synthesizer phase noise level and spurious tone specifications.

BT also requires from the frequency synthesizer the ability to generate communication channels (output frequencies) in 1MHz step (FCH). To perform these channels, the PLL should either have a 1MHz reference frequency, or have a fractional-N divider. In traditional fractional-N PLL, spurious level depends on loop bandwidth and sigma-delta order. Compensations of this type of spurious are possible, but are neither power nor area friendly.

We have designed an analog fractional-N PLL working at 2.45GHz with a Phase Switching Divider optimized for power consumption (about 1mW), see [1] for more details. The FPD PLL architecture produces no fractional division error, and that is energy-saving because it allows fractional PLL without sigma-delta, time-to-digital converter, or other averaging technique. This divider (Fig. 1) uses VCRO (Voltage Controlled Ring Oscillator) phases TPD's (Propagation delays) to generate the perfect fractional division (Fdiv) which is then compared to FREF (Reference Frequency) in the PFD (Phase and frequency detector).

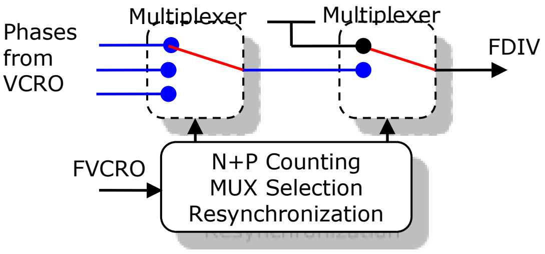


Fig.1: FPD principle bloc diagram

However this architecture of divider has poor spurious performance without calibration, because of inter-inverters propagation delay variations due to local process dispersions. This spurious have to be lowered below BT phase noise specifications: -90dBc@1MHz for the LO to be used for BT.

Thanks to the properties of FDSOI, the variability is lower in FDSOI than Bulk [2] and this calibration can be done acting on transistor's back gate voltages, adding no extra transistors in RF signal paths.

Three loops have been investigated [2] to ensure PLL operation over all Process Voltage and Temperature (PVT) corners:

1- A Frequency Locked Loop (FLL) corrects nominal frequency deviation.

2- A Duty Cycle Correction (DCC) balances Pmos/Nmos conductivity.

3- A multiplexed PLL loop unifies RO TPD's to reach a ±1% maximum deviation, corresponding to spurious level attenuation from -60dBc@1MHz to - 90dBc@1MHz.

Before implementing all PLL, 4 different VCO architectures have been sent to fabrication to verify Nominal Frequency, back gate Tuning, Pushing, Pulling and above all Phase noise specifications. Figure 2 shows the layout of the VCRO that will be used it the future PLL: a fifteen inverter Ring Oscillator.

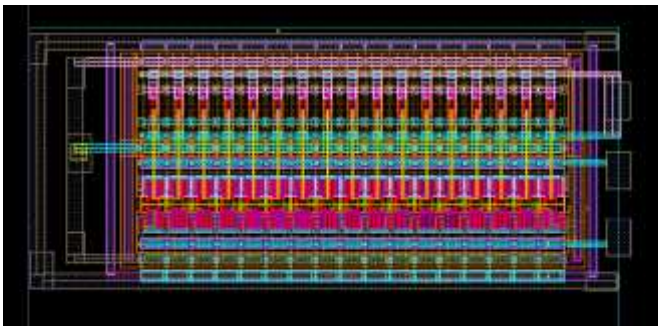


Fig.2: Fifteen inverters Ring Oscillator layout

Related Publications:
[1] A. Fonseca et al., "CMOS technology beyond 22 nm", International Conference on Small Science (ICSS 2013), Las Vegas, 15 to 18 December 2013.
[2] A. Fonseca et al., "Process variation compensation for PLL on FDSOI 28nm'' Proceedings of the VARI 2013, Karlsruhe, Germany
*et al. : E. De Foucauld (CEA-LETI), P. Lorenzini, G. Jacquemod (EPIB-UNS)

On-chip temperature monitoring for RF power amplifiers linearity test and self-healing

Research topics : CMOS RF integrated circuits, temperature sensor, testing, self-healing

J.L. González, J. Altet (UPC), D. Mateo (UPC) M. Vellevehi (CSIC-CNM), X. Jordà (CSIC-CNM)

ABSTRACT: This work aims at showing a new approach for determining the efficiency of linear class A RF power amplifiers by means of non-invasive, steady-state thermal monitoring. Silicon surface thermal monitoring is performed with built-in sensors and infrared measurements on an RF power amplifier. The first monitoring circuit consists of differential sensors, which can be used for contact-less on-line efficiency monitoring or for easy production testing. The obtained results are corroborated by means of infrared measurements.

The testing of RF and mmW integrated circuits is a complicated issue due to the high frequencies required and the complexity that this imposes to the tester. The same problems are encountered if self-healing strategies have to be implemented since the on-chip monitoring of high-frequency nodes of the circuit significantly degrades the circuit's performance due to the loading effect of the monitoring circuitry. However, there exist indirect observation methods that allow, at the same time, reducing the observation frequency and avoiding the contact to the high-frequency nodes. We have recently demonstrated that the DC power dissipated by the active devices of a power amplifier (PA) circuit carries information about its operation at high-frequencies [1].

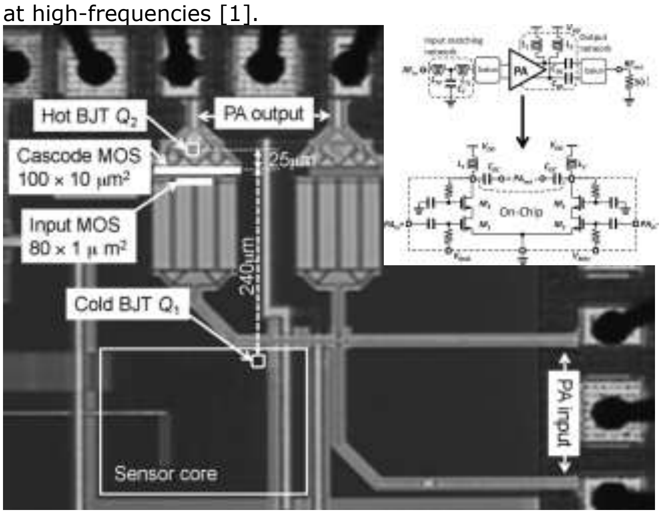


Figure 1 : Layout of the integrated PA and on-chip temp. sensor.

The temperature measured by on-chip differential sensors can effectively be used to determine the linearity and efficiency of RF PA [2,3]. Fig. 1 shows a prototype of a 2.5 GHz PA fabricated in a 65nm CMOS process along with an on-chip temperature sensor. The 'Hot BJT' transducer senses one of the PA active devices temperature without contacting it. The sensor provides a DC voltage that is proportional to the temperature difference between this transducer and the average surface temperature sensed by the 'Cold BJT', therefore making the sensor measurement independent to ambient temperature changes.

Fig. 2 illustrates how the sensor output can be used to determine the PA efficiency without requiring the complicated measurement of the PA output power at RF.

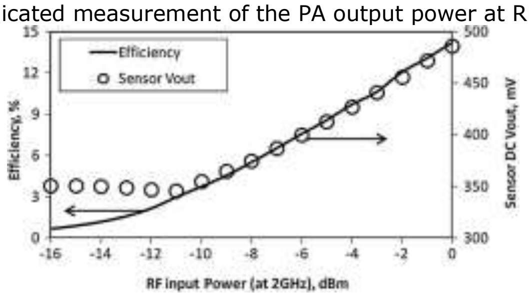


Figure 2: Comparison of electrically measured efficiency and information from the thermal sensor output.

The thermal monitoring technique by means of on-chip, contactless sensors has been validated in collaboration with the CSIC-CNM laboratory by comparing the temperature increases measured by the sensors and the results of an IR camera. The results of a differential temperature map are shown in Fig. 3, along with a comparison of the temperature measured by both methods.

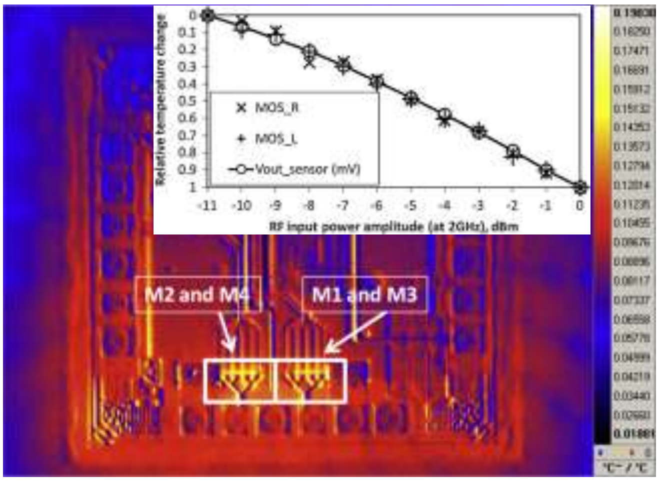


Figure 3: Validation of on-chip temperature measurements by off-chip IR thermography.

Related Publications:

[1] J. Altet, D. Gómez, C. Dufis, J.L. González, D. Mateo, X. Aragonés, F. Moll, A. Rubio, "On evaluating temperature as observable for CMOS technology variability" in Proc. of the 1st IEEE European workshop on CMOS Variability (VARI 2010).

[2] J. Altet, J. L. González, D. Gómez, X. Perpiñà, S. Grauby, C. Dufis, M. Vellvehi, D. Mateo, S. Dilhaire, X Jordà "Electro-thermal characterization of a differential temperature sensor and the thermal coupling in a 65nm CMOS IC", 18th Intl. Workshop on Thermal investigations of ICs and Systems, 25-27 Sep. 2012.

[3] J. Altet, D. Gomez, X. Perpinyà, D. Mateo, J. L. González, M. Vellvehi, X. Jordà, "Efficiency determination of RF linear power amplifiers by steady-state temperature monitoring using built-in sensors," Sensors and Actuators A: Physical, Volume 192, 1 April 2013, Pages 49-57.

High Efficiency Load Modulated RF Power Amplifier in SOI CMOS technology

Research topics : SOI, CMOS, PA, Efficiency, Tunable Matching, Load Modulation

A. Giry, G. Tant

ABSTRACT: This research work presents the implementation and measurement results of a SOI CMOS power amplifier with passive load modulation allowing enhanced efficiency at back-off power. The proposed PA architecture combines a SOI LDMOS device with a SOI CMOS tunable matching network based on high voltage switched capacitors. At 2.14GHz, the PA achieves 31.5dBm of measured peak power under 4V supply. Compared to a conventional PA with fixed matching, the measured efficiency improvement is 60 % at 8.5 dB back-off.

The challenge of high efficiency linear CMOS PA integration is exacerbated by the high bandwidth and high peak-to-average power ratio (PAPR) of LTE signals. Conventional linear PAs make use of a fixed output matching network designed to maximize efficiency at peak power, and efficiency degradation occurs when lowering power. With high PAPR modulated signals, the PA has to operate in back-off most of the time, resulting in relatively low average efficiency. Passive load modulation is a promising technique which aims at PA efficiency optimization at back-off power by providing load impedance modulation to the PA thanks to a tunable matching network. Compared to other high-efficiency PA architectures (Doherty, Envelope Tracking…), passive load modulation provides an attractive trade-off among performance, cost and integration capability and is well suited for CMOS integration.

In order to improve PA efficiency in back-off at reduced cost and size, passive load modulation of a SOI PA has been investigated. Today, SOI CMOS technology offers new opportunities for PA integration [1]. The use of a high resistivity substrate enables transistor stacking which allows the design of high power RF switch, a crucial feature for the implementation of low loss tunable matching networks (TMN). In this work, a SOI LDMOS PA and a SOI CMOS TMN using high voltage switched capacitors have been designed and implemented. The proposed load modulated PA [2,3] consists of a SOI LDMOS device followed by a fixed output pre-matching network and a SOI CMOS TMN as shown in Fig.1.

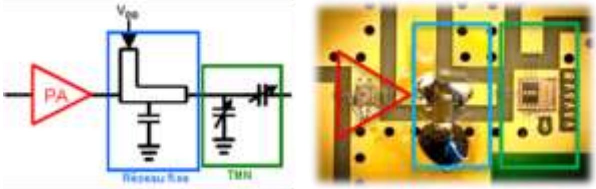


Figure 1 : Architecture and micrograph of the proposed SOI CMOS load modulated PA

The load impedance of the SOI LDMOS power device can be tuned thanks to the TMN which is integrated in the same SOI CMOS process and occupies an area of 1.7mm2. A co-design approach has been adopted in order to optimize the RF performances of the whole PA. The TMN includes series and shunt tunable capacitors, both designed so that the TMN covers the optimal PA load impedance trajectory as shown in Fig.2. Tunable capacitors consist of arrays of binary weighted switched capacitors using stacked floating body transistors as switches. In order to minimize insertion loss of the TMN, the NMOS transistors were sized to obtain a minimum quality factor of 50.

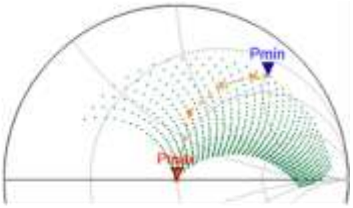


Figure 2 : Impedances coverage (green dots) of the TMN and optimal load trajectory (orange dashed line)

Fig.3 shows the simulated efficiency of the PA with (green dots) or without (blue curve) load modulation, the red curve corresponding to the optimal states. A maximum efficiency of 65% is obtained at peak power. With fixed matching, efficiency decreases to 30% at 7.5dB back-off, whereas load modulation provides 45% efficiency despite TMN loss.

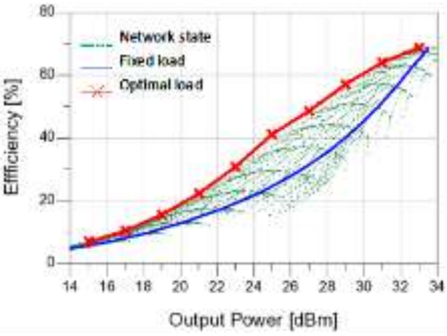


Figure 3 : Efficiency of the load modulated PA as a function of output power (red curve)

From 6 to 12 dB back-off, more than 50% efficiency improvement has been measured at 2.14GHz on the realized prototype.

Related Publications :
[1] Giry, A.; Tant, G.; Lamy, Y.; Raynaud, C.; Vincent, P.; Bertrand, G.; Joblot, S.; Velard, R.; Coudrain, P.; Carpentier, J.; Petit, D. & Rauber, B., "A monolithic watt-level SOI LDMOS linear power amplifier with through silicon via for 4G cellular applications'', 7th IEEE Radio and Wireless Week (RWW), 2013
[2] Tant, G.; Giry, A.; Arnould, J.-D.; Fournier, J.-M. & Vincent, P., "Amplificateur de puissance à réseau de charge accordable en technologie SOI CMOS pour station de base Femtocell 4G'', Proceedings of the Journées Nationales Microondes 2013
[3] Tant, G.; Giry, A.; Vincent, P.; Arnould, J.-D. & Fournier, J.-M., "A 2.14GHz watt-level power amplifier with passive load modulation in a SOI CMOS technology'', Proceedings of the ESSCIRC, 2013

Offering a soul to inanimate objects

Research topics: Coordination language, rule-base, distributed systems

H. Iris, M Louvel, F Pacull, N. Geraud (Dasein - Interactions)

ABSTRACT: Alphonse de Lamartine was wondering "Inanimate objects, do you have a soul ?". A couple of centuries later the question can be raised again in the frame of the Internet of Things. We relate here after two contributions in this domain based on our in-house LINC middleware. First, thanks to a table embedding several hundreds of RFID readers and tagged physical objects (3D printed buildings) we associate a physical model of urban scene to its virtual 3D representation. In the second contribution we bring some smartness to the classical sensors and actuators in order to allow them to be first citizen in complex coordination.

Moving objects from the status of "inanimate" to "first citizen" of their world was a dream for poets of the previous centuries. Step by step it becomes a reality in the frame of the so-called Internet-of-Things. In the following we describe two of our contributions in the subject.

The first example [1] is in the context of a very innovative hardware table stacking a 48 inches HD screen on top of a matrix of 384 RFID readers. This table, along with tagged 3D printed buildings, actions cards, tokens, tangible objects, and 3D mouse are used as inputs for urban mediation. It allows mapping the physical model of the urban scene to its virtual representation (Figure-1).

For instance, it is possible to move a token representing a pedestrian within the physical model and acting on her field of vision (moving her head) via a 3D mouse. As a result we can see on an external screen the 3D virtual scene what the pedestrian actually sees. Modifications of the physical model are imme-diately echoed on the 3D virtual scene and thus, the potential impact can be better evaluated. This table may be used to capture the viewpoints of the different involved people: current residents, potential inhabitants of the future buildings and the city or area managers responsible of the project.

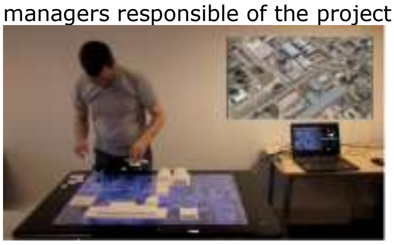


Figure 3 - Urban Mediation

The middleware LINC has the task to synchronize a huge set of inputs coming from the matrix of RFID readers, a 3D mouse and other connected interfaces. The power of the coordination model of this middleware simplifies a lot the design of the rules responsible for tracking the interactions, the feedback on the table screen (background) and on the 3D engine displaying the scene.

The resource based approach of the middleware offers a high level abstraction layer which allows interfacing through the same paradigm components that, at a first thought we could consider as very different, e.g. a RFID reader and a 3D engine. However, the LINC approach allows designing such complex applications with only a couple of coordination rules.

The full system is modeled as bags containing resources describing the current status and acting on it. Indeed, coordination rules waiting for the presence of some resources in some bags are responsible for the evolution of the system by the insertion and/or removal of resources.

These actions on the resources make the system evolve.

Both observers and modifiers of the system rely on the same bag paradigm.

The second contribution was to raise the classical sensors and actuators that are usually either autonomous and/or passive to the rank of first citizen of the middleware by allowing them to take effective part of the coordination.

This brings two major interests.

First, this concerns the possibility for the sensors and the actuators to get, as part on the coordination protocol, inputs from the outside (the coordination in the context of an application). This may include for instance receiving the information of wake-up from the application by a low-power mechanism that is most of the time not associated to the regular RF communication circuit. During the sleep phase, a sensor can be put in very low power consumption mode waiting only for this wake-up signal. For instance, let's consider a flood monitoring systems [2] based on sensors that are positioned along the banks of a river. The first sensor is located upstream and embeds a mechanical mechanism able to generate a signal when the latter is touched by the water. The mechanism consists of a compressed spring trapped in a water-soluble matrix. When the matrix is altered by the water, the spring is released producing enough energy to emit the signal. The other sensors distributed along the banks are sleeping, waiting for the wake-up signal sent by the application as preamble of the coordination protocol.

Second, with the coordination protocol we let the possibility to the sensor or actuator to express the fact they will not be able to do what the coordination is waiting for. They just declare in the first phase of a transactional process that they are not reading the awaited value of a signal or the actuation currently requested is not feasible (lack of energy, out of functional range …). Thus, the full transaction is aborted avoiding the always embarrassing situation where only part of the job is done and the system is globally inconsistent. The coordination protocol at the device level is so compact that is can be implemented on same very small micro-controller usually used for sensors and actuators. This simplifies a lot the design of the application. For instance, we manage the obstacle avoidance of a robot with a single rule.

Related Publications:
[1] M. Louvel and F. Pacull. A coordinated matrix of RFID readers as interactions input. In SENSORDEVICES 2013, The Fourth International Conference on Sensor Device Technologies and Applications, pages 91–96, 2013.
[2] H. Iris and F. Pacull. Smart sensors and actuators: A question of discipline. Sensors & Transducers Journal, 18 (special Issue jan 2013):14–23, 2013.

Self-aware cyber-physical systems and applications in smart buildings and cities

Research topics : Cyber-physical systems, Autonomic computing, Self-aware systems

Levent GURGEN, Ozan GUNALP, Yazid BENAZZOUZ, Mathieu GALLISSOT

ICT has a substantial potential to help cities to respond to the growing demands of more efficient, sustainable, and increased quality of life in the cities, thus to make them "smarter". Smartness is directly proportional to the "awareness". Cyber-physical systems can extract the awareness information from the physical world and process this information in the cyber-world. Thus, a holistic approach, from the physical to the cyber-world is necessary for a successful and sustainable smart city. This research work introduces important challenges and provides guidelines and recommendations to achieve self-aware smart city objectives.

Currently more than half of the world population lives in cities and the urban areas of the world are expected to absorb all the population growth expected over the next four decades while at the same time drawing in some of the rural population. Besides, on 2% of the earth's surface, cities actually use 75% of the world resources. These facts make naturally the cities important actors for the world's sustainable development strategy. One immediate action of the governments in the world has been to take measures in order to transform cities into "smart cities" that better manage their resources. ICT has a substantial potential to help cities to respond to the growing demands of more efficient, sustainable, and increased quality of life, thus to make them "smarter" and context-aware.

Existing ICT solutions do not provide the support required for applications to cope with a dynamically changing physical context. On the other hand, Cyber-physical systems form the interface between the physical real world and the cyber world (see Figure 1), should thus bring the necessary mechanisms and tools that would make applications aware of the changes in the physical context and adapt their execution according to it. By building such cyber-physical systems, an important step will be taken towards building the city nervous system that would provide the awareness to the city: the Smart City Ecosystem.

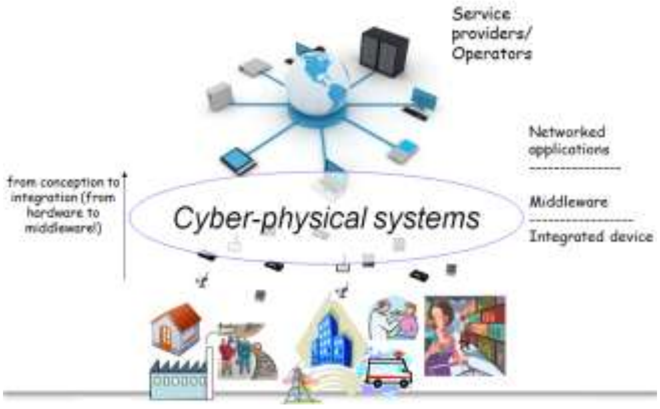

Service providers/ Operators
Networked applications
Middleware
Integrated device
from conception to integration (from hardware to middleware!)
Cyber-physical systems

Figure 1 : Convergence of the physical and the virtual world

We propose to build such "self-aware cyber-physical systems with the MAPE-K model, which basically consists of a control loop with 4 phases - Monitor, Analyze, Plan, Execute - in continuous interaction with a Knowledge base (See Figure 2). The monitoring phase concerns continuous monitoring of different properties of managed elements (e.g., CPS devices such as sensors, actuators, gateways; CPS services; physical environment). In the Analysis phase, collected "raw" data are processed and analyzed in order to obtain valuable information on the state of the managed elements. Given the situation and context information obtained from the analysis phase, the planning phase gives decisions for actions to take in order to attain the high-level objectives defined by humans. According to the decisions taken in the planning phase, the execution phase schedules and executes the decided actions on the managed element.

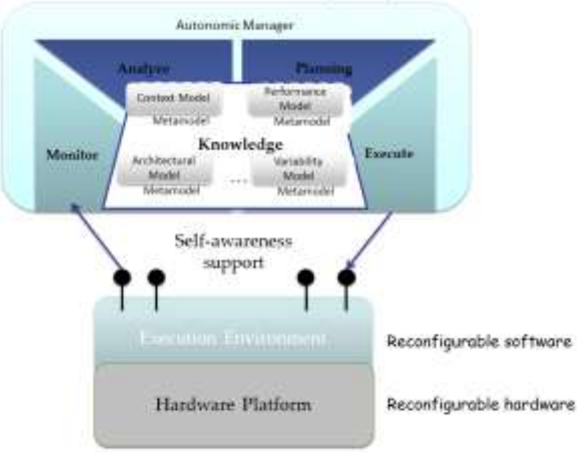


Figure 2 : Self-awareness: monitor, analyze, plan, execute

Applying the MAPE-K model in cyber-physical systems necessitates adequate computing infrastructures. Service-oriented architecture (SOA) is one candidate technology that offers primitives for creating modular, reconfigurable and extensible software platforms. Service-based approach allows composing flexible, robust and adaptable applications from platform services and base services developed by different city stakeholders such as telecom operators, service providers or device manufacturers.

The coordination of sensors and actuators is thus handled using a service composition approach. Devices export their functionalities in terms of services. Application creation thus consists of composing these services to obtain higher level composed services. We provide tools to assist developers to create with a model-based approach to guarantee robustness, correctness and self-adaptation properties.

Related Publications:
[1] L. Gurgen, O. Gunalp, Y. Benazzouz, M. Gallissot, "Self-aware cyber-physical systems and applications in smart buildings and cities", International Conference on design, Automation & Test in Europe (DATE), 2013.

Coordination among Building Automation Systems

Research topics: distributed system, smart building, coordination and control

F Pacull, S. Lesecq, O. Yaakoubi, S. Thior, L.F. Ducreux, C. Guyon-Gardeux, H. Moner (UTRC-I), D. Pesch (CIT), A. McGibney (CIT), F. Bernier (Schneider) F. bonnard (Schneider)

ABSTRACT: Nowadays, Building Automation can be considered a particular case of Network controlled systems as each Building Automation System possesses its own network in order to deliver various kinds of information to the different devices. Networks introduce drawbacks such as delays or data loss that must be taken into account during the design phase in order to ensure the required level of performance for the control and monitoring.

Building Automation can be seen as a particular case of Network Controlled Systems (NCS): a network (wired and/or wireless) is present in the system in order to deliver information from the sensors to the controllers/monitors then to the actuators. This subsystem introduces drawbacks such as delays, data loss, etc. that must be explicitly taken into account at the control level in order to ensure the required "Quality-of-control" [1].
The control of such NCS is eased by the integration of the LINC coordination middleware. This latter offers an abstraction layer that hides all the technical details regarding the underlying technologies [2][3]. In this way, the user can concentrate its effort on the functionality (e.g. a particular control) he/she has to implement, whatever the technologies of the different devices is. As a consequence, cooperation among different subsystems can be done without the integration of costly hardware gateways that "translate" a particular protocol to another one in order to make heterogeneous technology exchange information.

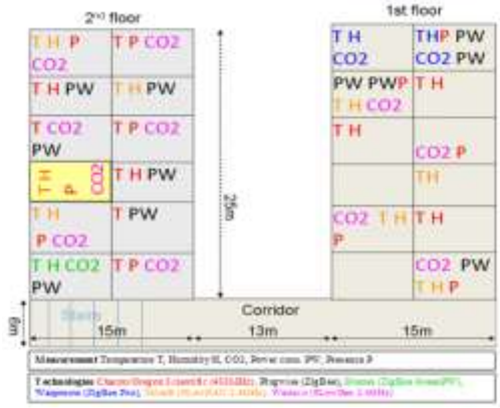


Figure 1: Heterogeneous sensor deployment over the CTL building

Moreover, the abstraction layer offers the capability to create complex scenarios because all data are seen in a unified way [4]. As a consequence, monitoring and control of a smart building can be done taking into account all the devices (in different technologies) deployed in the building (see Fig. 1 where the monitoring of the CTL building is performed with 7 different wireless technologies) [7].

In the SCUBA FP7 project (nb 288079), the LINC middleware is the architecture backbone (Fig. 2). Various technologies can be seen in a unified way (see lower part of Fig. 2) thanks to the abstraction layer it introduces. Moreover, it offers coordination among various systems and components in the tool and middleware layers (e.g. BASOnt, Rescue Worker Interface, Strategic Manager) [5][6].

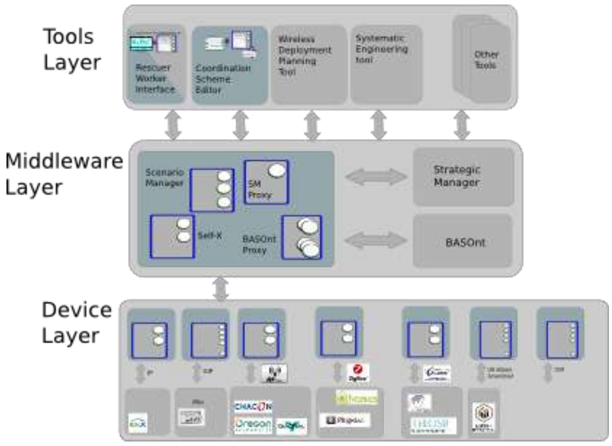


Figure 2: SCUBA architecture. The LINC middleware (Self-X) ensures the coordination among various Building Automation subsystems. It provides an abstraction layer that hides the technology heterogeneity.

Related Publications:
[1] S. Lesecq, Control over a network. Context and Challenges, CERIST Autumn School on Cyber Physical Systems, Algier, Algeria, Sept 30 - Oct 3, 2013.
[2] F. Pacull, LINC: Coordination Middleware from legacy services to sensor-actuator networks, CERIST Autumn School on Cyber Physical Systems, Algier, Algeria, Sept 30 - Oct 3, 2013.
[3] F. Pacull, LINC middleware basics, IECON Tutorial, Vienna, Austria, Nov 2013.
[4] F. Pacull, Creating complex heterogeneous scenarios with the Coordination Scheme Editor, IECON Tutorial, Vienna, Austria, Nov 2013.
[5] F. Pacull et al., Self-organisation for Building Automation Systems: Middleware LINC as an Integration Tool, IECON Special session, Vienna, Austria, Nov. 2013.
[6] F. Bernier et al., Architecture for Self-organizing, Cooperative and Robust Building Automation Systems, IECON Special session, Vienna, austria, Nov. 2013.
[7] A. McGibney et al., Wireless Sensor Networks for Building Monitoring : Deployment Challenges, Tools and Experience, Fifth Workshop on Real-World Wireless Sensor Networks (REALWSN 2013), 19-20 Sept. 2013, Como lake, Italy.

A first approach of diagnosis strategy for complex wired networks

Research topics: cable, reflectometry, Bayesian networks, communication

W.BEN HASSEN, F.AUZANNEAU, F.PERES and A. TCHANGANI (LGP, ENIT, INPT)

ABSTRACT: Although reflectometry method has proven its efficiency in locating faults for simple topologies, it still suffers from ambiguity problems in the case of branched networks. Distributing several diagnosis systems is a possible answer, but this raises new problems, which are related to "diagnosis strategy": how to optimize the number and position of sensors to maximize the diagnosis coverage without decreasing performances. Here, a Bayesian Network model for diagnosis shows how several sensors can work together to improve the diagnosis of a complex network and how communication between sensors helps to reduce uncertainties.

In this work, we present a Bayesian Network (BN) approach for the diagnosis of branched networks. Our main objective is to find a good compromise between the system cost (i.e. the diagnosis systems number) and the overall diagnosis performance.

In order to reduce the number of sensors, the wiring network is divided into several sub-networks where each one is diagnosed by only one diagnosis sensor as shown in Fig.1.

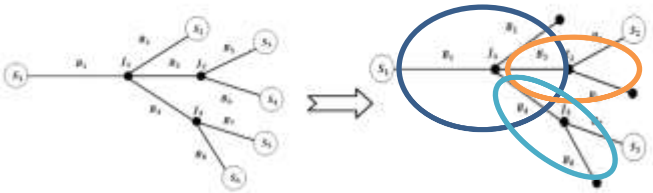


Figure 1: Sensors optimization (left: global case, right: sub-network case)

The proposed approach is described schematically on Fig. 2

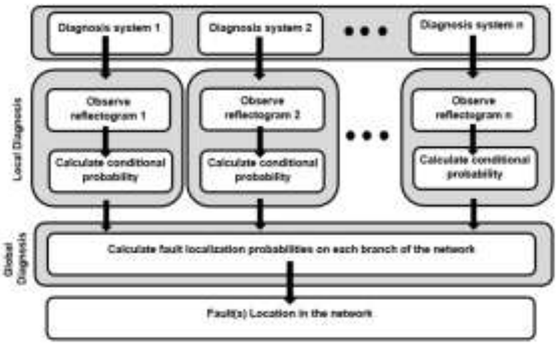

Diagnosis system 1
Diagnosis system 2
Diagnosis system n
Local Diagnosis
Observe reflectogram 1
Observe reflectogram 2
Observe reflectogram n
Calculate conditional probability
Calculate conditional probability
Calculate conditional probability
Global Diagnosis
Calculate fault localization probabilities on each branch of the network
Fault(s) Location in the network

Figure 2: Diagnosis Procedure in Wired Networks

To reduce complexity, the BN is divided into multiple sub-systems where each one models the local diagnosis as shown on Fig.3. The local BN explicitly estimates the health of the diagnosis system and the status of the wires.

Then, local BNs are integrated into a global BN in order to locate faults in the whole network [1,2].

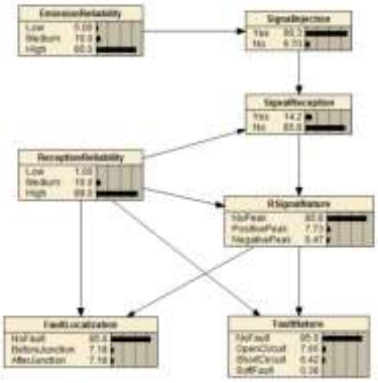


Figure 3: Elemental Bayesian Network

A fault is simulated on branch B3. In this case, even if a low system cost is achieved (from 6 sensors down to 3), the obtained diagnosis quality is low (confidence level equal to 33%). As a solution, communication between neighboring sensors is introduced to exchange information about the detected fault [3]. Fig.4 shows the global Bayesian network in the optimized case. Here, both, diagnosis quality and confidence level are high.

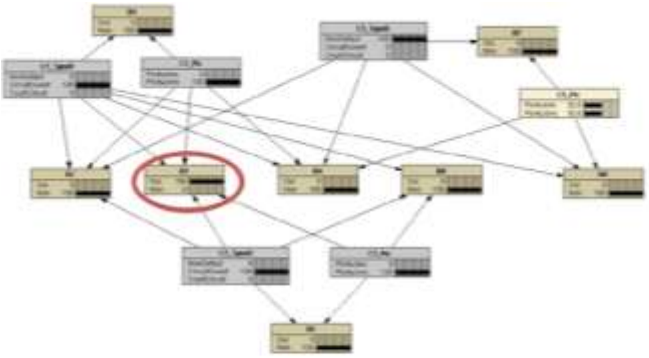


Figure 4: The Global BN with sensors communication

In future works, the life profile of the cable (Environment, cable characteristic) will be introduced for more optimization.

Related Publications:
[1] W. B. HASSEN, F. AUZANNEAU, F. PERES, and A. TCHANGANI, "A Distributed Diagnosis Strategy using Bayesian Network for Complex Wiring Networks," in IFAC Workshop on Advanced Maintenance Engineering, Services and Technology (AMEST), November 2012.
[2] W. B. HASSEN, F. AUZANNEAU, F. PERES, and A. TCHANGANI, "Optimisation de Capteurs de Diagnostic de Défauts par Réflectométrie dans les Réseaux Filaires Complexes en utilisant les Réseaux Bayésiens, " in Congrès International Pluridisciplinaire Qualité et Sûreté de Fonctionnement (Qualita), March 2013.
[3] W. B. HASSEN, F. AUZANNEAU, L. INCARBONE, F. PERES, and A. TCHANGANI, "OMTDR using BER Estimation for Ambiguities Cancellation in Ramified Networks Diagnosis, " in IEEE ISSNIP, " April 2013.

OMTDR using Communication and BER estimation for Branched Networks Diagnosis

Research topics: wiring network, diagnosis, communication, OMTDR, BER, sensor fusion

W.BEN HASSEN, F. AUZANNEAU, L. INCARBONE, F.PERES and A. TCHANGANI (LGP, ENIT, INPT)

ABSTRACT: A new reflectometry method, named "Orthogonal Multi-Tone Time Domain Reflectometry" (OMTDR) is proposed. Based on the application of OFDM communication method, OMTDR is applied to on-line diagnosis of complex wired networks. OMTDR uses the transmitted part of the test signal to enable reflectometry systems communication, avoiding ambiguities related to fault location in branched networks. Taking advantage of Bit Error Rate (BER) estimation additionally helps locating and evaluating the severity of soft defects.

On-line diagnosis consists in detecting and locating defects while the system is operating. Since critical signals are present on the network, reflectometry-based on-line diagnosis imposes serious challenges on harmlessness, bandwidth control and interference mitigation. Moreover, in the case of complex branched networks, using a single reflectometry system suffers from an ambiguity problem. As a solution, a distributed diagnosis strategy is applied. It consists in making reflectometry measurements at several ends of the network to get different perspectives. This requires that the signal from one diagnosis system does not interfere with the others.

In this context, a new method, called "Orthogonal Multi-Tone Time Domain Reflectometry" (OMTDR), is proposed. Based on Orthogonal Frequency Division Multiplexing (OFDM) method, OMTDR permits interference avoidance, complete bandwidth control and spectral efficiency thanks to the orthogonality among tones [1].

In a branched network, using a single reflectometry sensor may result in ambiguous location on defects. This may be solved by placing several sensors at several ends of the network. OMTDR proposes to add communication to the diagnosis function of the sensors. Communication not only permits to evaluate the channel state [2], but also to exchange information about the defect's location. Then, the aggregation of all this information provides unambiguous defect's location [3].

Using TDMA inspired strategy (Fig. 1) enables to ensure that the chosen master reflectometers will properly gather all information from its slaves (sensor fusion).

Communication offers an additional advantage: the estimation of the Signal to Noise Ratio (SNR) and Bit Error Rate (BER). We have shown that the presence of a defect on a wire on the path from a sensor to another one increases the BER, according to its severity (Fig. 2). BER measurement between reflectometers provides information about the channel state. Then measuring BER values while doing reflectometry not only permits to pre-locate the detected defect, but also to evaluate its severity.

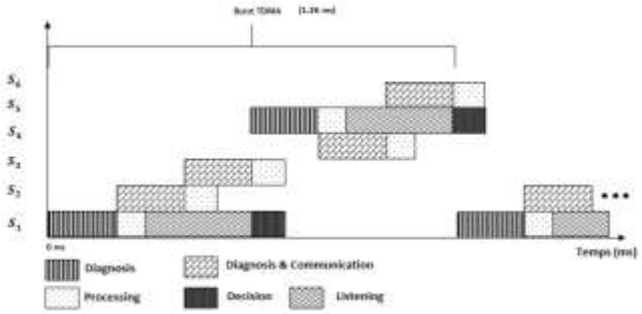


Figure 1: TDMA-like bandwidth allocation for several OMTDR sensors.

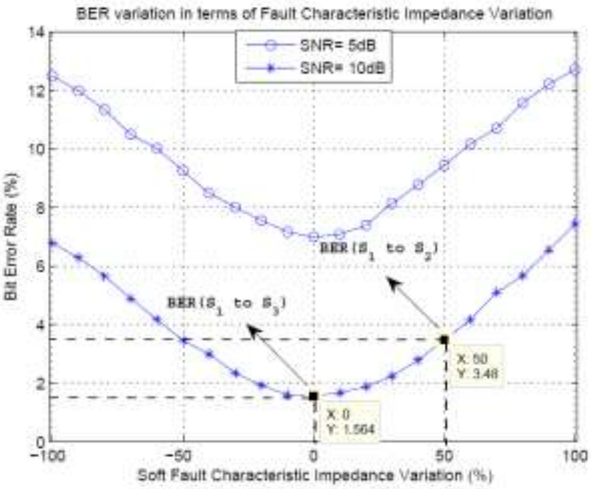


Figure 2: Defect's pre-location using BER estimation.

Fig.2 shows that the BER value is higher for communication between sensor S1 and S2 than between sensors S1 and S3: the defect is on the branch of the path S1-S2 than does not belong to the path S1-S3. Its severity may be estimated to a local variation of 50% of the characteristic impedance of the wire. The precise location is then deduced from standard reflectometry analysis.

Related Publications:
[1] W. B. HASSEN, F. AUZANNEAU, L. INCARBONE, F. PERES, and A. TCHANGANI, "OMTDR using BER Estimation for Ambiguities Cancellation in Ramified Networks Diagnosis", in IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing (IEEE ISSNIP'13)", April 2013.
[2] W. B. HASSEN, F. AUZANNEAU, L. INCARBONE, F. PERES, and A. TCHANGANI. "On-Line Diagnosis using Orthogonal Multi-Tone Time Domain Reflectometry in a Lossy Cable", in IEEE Multi-Conference on Signals, Systems and Devices (IEEE SSD'13), March, 2013.
[3] W. B. HASSEN, F. AUZANNEAU, F. PERES, and A. TCHANGAN. Diagnosis Sensor Fusion for Wire Fault Location in CAN Bus Systems. IEEE SENSORS 2013, November, 2013.

New methods to improve the diagnosis of soft defects in cables

Research topics: wire diagnosis, soft defects, time reversal

L. El Sahmarany, L. Incarbone, F. Auzanneau

ABSTRACT: Reflectometry-based methods are widely used for the detection and location of defects in wired networks. They provide accurate position of hard defects (open and short circuits) but have limited performances in case of very long wires or soft defects (chafing, corrosion, etc.). A new post-processing method helps improve these results and enable to increase the maximum distance of defect detection and soft defect detection capacity. Additionally, the application of time reversal techniques provides a new way for incipient or soft defects detection and location.

The need for detection and location of defects in wired networks has been recognized in many application domains such as power distribution, communications and transportation. Several methods have been investigated such as impedance spectroscopy, infrared thermography or X-rays imagery but the most promising ones are based on reflectometry [1]. A high frequency signal is injected at one end of a wire and propagates through the network. Each time an impedance discontinuity (branch, connector or defect) is met, a part of the signal's energy is sent back. The analysis of the reflected signal provides information about faults characterization and location.

These methods provide very good results for hard faults detection, i.e. open and short circuits. Location accuracy is a few percent of the total length of the wire. But this result may be decreased in the cases of very long wires and soft faults detection (chafing, corrosion, hot points, etc.). The need for soft faults diagnosis has led to the development of new methods, more adapted to these problems.

Two different methods have been proposed. The first one is a model-based approach, using a specific post-processing procedure [2] aiming at reducing the effects of dispersion, which tends to enlarge the signal throughout its propagation in the wire and decrease the diagnosis performances (detection capacity and location accuracy).

Usual processing of the measured signal uses cross-correlation between the injected and the reflected signals. This operation measures the similarity of the reflected waveform y and the injected one x as a function of the time-lag θ applied between them. This does not take the propagation medium into account: a constant velocity v is assumed, so that the time measurements can be converted into distances. The new method uses a model-based varying reference to better consider the effects of the wire on the propagating signal. This changes the usual cross-correlation formula

$$R_{xy}(t) = \int_{-\infty}^{+\infty} x(\theta) y(t+\theta) d\theta$$

to the following

$$R'_{xy}(t) = \int_{-\infty}^{+\infty} \frac{1}{A_\theta} \int_{-\infty}^{+\infty} X_0(\omega) e^{-\gamma(\omega)v(\omega_0)\theta}\, e^{j\omega t}\, d\omega\, y(t+\theta) d\theta$$

where X_0 is the spectrum of the injected signal and $\gamma(\omega)$ is the propagation constant of the cable.
This operation is called "dynamic correlation" because the reference signal changes for each value of the time-lag θ.

Figure 1 shows that dynamic correlation increases the detection capacity (green arrow), improves the location accuracy for long length wires and better noise immunity.

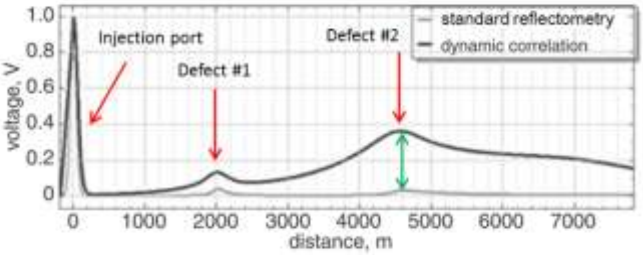


Figure 1: Performance improvement using dynamic correlation

The second method takes advantage of the focalization property of Time Reversal to improve soft defect's detection. A standard reflectometry process is done in a new cable and in a faulty one. The measured signals are used to generate the voltage maps in these 2 cables, which are convoluted to the input signal. The comparison of these results drastically amplifies the peaks of the defects, as shown on Figure 2.

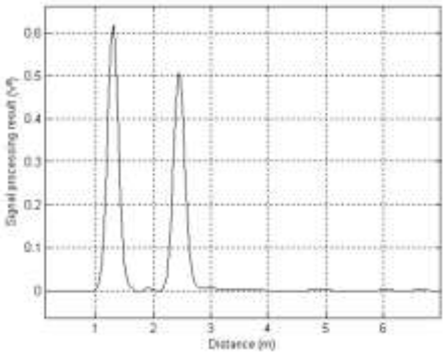


Figure 2: Detection of 2 soft defects in a 4m long coaxial cable

The signature of these defects using standard reflectometry would be close to the noise level: the new method greatly amplifies the peaks of the defects, easing their detection.

Related Publications:
[1] F. Auzanneau, "Wire Troubleshooting and Diagnosis: Review and Perspectives", Progress in Electromagnetics Research B 49, 253-273
[2] L. Sommervogel, L. El Sahmarany, L. Incarbone, "A method to compensate dispersion effect applied to Time Domain Reflectometry", Electronics Letters 49(18), 1154-1155, 2013
[3] L. El Sahmarany, N. Ravot, F. Auzanneau, P. Bonnet (LASMEA), L. Berry (LASMEA), "Time Reversal for Soft Fault Diagnosis in Wire Networks", Progress in Electromagnetics Research (PIER) 31, 45-58, 2013

2

Digital Architectures & Systems

3D ICs

Many Cores

Reliability

Real Time OS

Code Parallelization

Compilers

Design Methodologies

Adaptive cache architecture exploiting 3D stacking

Research topics: cache, 3D TSV, manycore

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

ABSTRACT: The number of cores in a Massively Parallel System on Chip (MPSoC) directly drives the needs of memory bandwidth, resulting in the well-known memory wall issue. 3D technologies allow the stacking of memory on top of the processing tier, thus allowing to build big 3D caches or to embed the main memory on chip. This work proposes an adaptive 3D cache architecture and evaluates its performances. A 1 MB cache tile has an area of 1.5 mm² in 28nm bulk LP technology and consumes 130.8 mW while providing 56,5 Gb/s of available bandwidth.

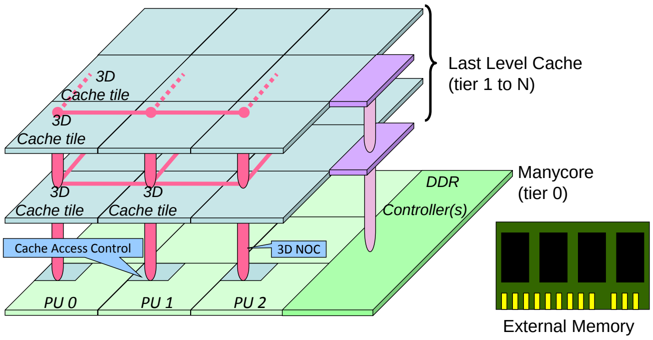


Figure 1: 3D adaptive cache architecture for manycores

The number of cores in a Massively Parallel System on Chip (MPSoC) directly drives the needs of memory bandwidth, resulting in the well-known memory wall issue. As the number of processing elements grows, the pressure on the external main memory can be alleviated through the use of big embedded caches. 3D technologies allow the stacking of memory on top of the processing tier, thus allowing to build big 3D caches or to embed the main memory on chip.

We propose a 3D non uniform cache architecture optimized for manycore environments with a high degree of parallelism as shown in Fig. 1. 3D cache tiles, which are autonomous caches, are stacked on top of the manycore architecture. These cache tiles are interconnected through a 3D NoC. This 3D NoC also connects them to cache access controllers located in the manycore tier. The cache access controllers are responsible for distributing memory accesses to cache tiles and locating the data in the 3D stack.

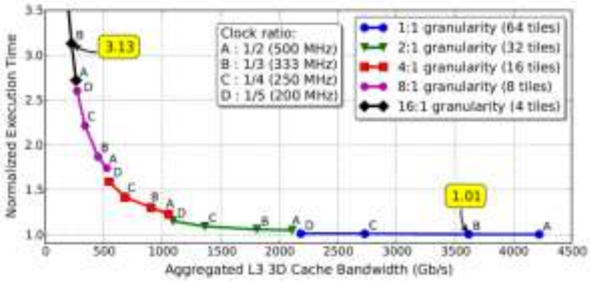
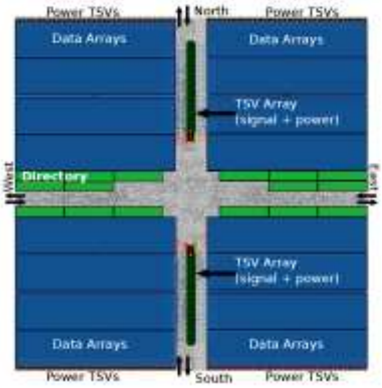


Figure 2: FFT execution time in function of 3D cache offered bandwidth

Even if using bigger cache tiles is more efficient for area utilization, for a given cache size the total number of cache tiles is reduced. Therefore, having less cache tiles results in a lower aggregated cache bandwidth. So, we searched the best cache configuration for a manycore usage in order to find a trade-off between performance, power consumption and area. Fig. 2 shows the evolution of execution time of an FFT running on a 256 cores platform in function of the number of the offered bandwidth of our 3D cache used as a L3 cache. The granularity of our architecture and the frequency ratio of the cache tiles according to the 3D NoC modulate this bandwidth. The finest granularity with a reduced cache tile's clock frequency provides the best performance/area trade-off.



Technology	3D ST CMOS28LP
TSVs	10 µm pitch
Microbumps	20 µm pitch
3D signals	284 signals/280 power
Cache freq.	390 MHz (typ)
3D NoC freq.	1 GHz (typ)
Total area	1225x1225 µm
Data area	77%
Peak BW	56.5 Gb/s
Power eff.	2.31 pJ/bit

Figure 3: Physical implementation of a 1 MB 3D cache tile

We implemented a 1 MB 3D cache tile in 28nm bulk 3D technology as shown in Fig. 3. The number of power TSVs is dimensioned to support up to 4 tiers of 3D cache. Our architecture provides a high storage efficiency with 77% of the 1.5 mm² occupied by the data of the cache. The 5µm-wide TSVs represent 4% of the total area. Conversely, 10µm TSVs (40µm pitch-similar to WIDEIO implementation) would take up to 33% of the tile area, showing that high density TSVs are mandatory for distributed caches application. Moreover, the power efficiency of the tile (2.31 pJ/bit) is better compared to Wide IO 3D memories (2.8 pJ/bit). During the execution of real High Performance Computing (HPC) applications, the 3D cache consumes about 23 mW per tile of 1 MB. So a big cache of 256 MB on 4 tiers consumes 5.9W while providing an aggregated bandwidth of 6 Tb/s.

Related Publications:
[1] Guthmuller, E.; Miro-Panades, I. & Greiner, A. (2013), 'Architectural exploration of a fine-grained 3D cache for high performance in a manycore context'', IFIP/IEEE 21st International Conference on Very Large Scale Integration (VLSI-SoC), 2013, 302-307
[2] Dutoit, D.; Guthmuller, E. & Miro-Panades, I. (2013), '3D integration for power-efficient computing'', 16th Design, Automation and Test in Europe Conference and Exhibition, DATE 2013, 18-22 March 2013, Grenoble', 779-784.
[3] Clermidy, F.; Dutoit, D.; Guthmuller, E.; Miro-Panades, I. & Vivet, P. (2013), '3D stacking for multi-core architectures: From WIDEIO to distributed caches', 2013 IEEE International Symposium on Circuits and Systems, ISCAS 2013, Beijing', 537-540.

A 0.9 pJ/bit, 12.8 GByte/s Wide-I/O Memory Interface in a 3D-IC NoC-based MPSoC.

Research topics : 3D Integrated Circuits, WideIO SDRAM, MPSoC, Network-on-Chip.

D. Dutoit, C. Bernard, S. Chéramy, F. Clermidy, Y. Thonnart, P. Vivet, C. Freund*, V. Guérin*, S. Guilhot*, S. Lecomte*, G. Qualizza*, J. Pruvost, Y. Dodo**, N. Hotelier**, J. Michailos**. (*STE), (**ST)**

ABSTRACT: 3D Integrated Circuit (3D-IC) opens architecture opportunities for improved SoC-to-memory interconnect bandwidth between dies. This research topic consists in the design of a two-tier 3D-IC composed of one NoC-based MPSoC and one multi-channel WideIO mobile SDRAM stacked in a face-to-back configuration. Measurements of the 3D-IC show that the targeted 12.8GByte/s bandwidth is achieved in worst case conditions, while offering a 0.9 pJ/bit 3D I/O link power efficiency.

Three-dimensional (3D) stacking technology, which enables low-latency, high bandwidth and very dense die-to-die interconnects, is the right opportunity to develop new vertical connection schemes for memory chip stacked upon a MPSoC.
Within an existing MPSoC architecture organized around a 16-router Asynchronous NoC backbone, we have embedded [1] four identical advanced memory traffic managers (figure 1) to interface with each channel of the new standardized WideIO mobile DRAM (JEDEC: http://www.jedec.org). The PHYsical layer consists in a 128-bit wide DRAM data capture interface and targets operations up to 200 MHz in Single Data Rate (SDR) mode, offering an aggregate 12.8 GB/s memory link. A simple digital std-cell based design in favor of ultra-low power consumption by removing all power hungry programmable Delay-Locked Loop devices (DLL) has been achieved.

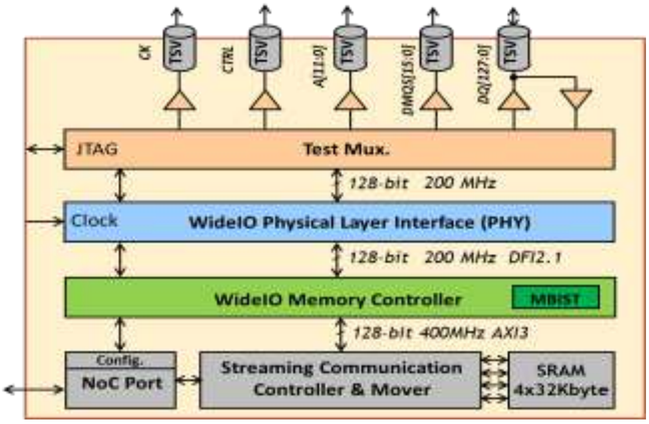


Figure 1 : WideIO interface architecture

The MPSoC has been implemented in STMicroelectronics' low-power 65nm CMOS process and assembled with a WideIO memory stacked in a Face-to-Back configuration (figure 2). The interconnect, located in the center area of the memory die, consists of an array of micro-bumps with a 40 µm x 50 µm pitch. With this configuration, TSVs are required within the MPSoC for WideIO memory supply distribution and signal propagation.

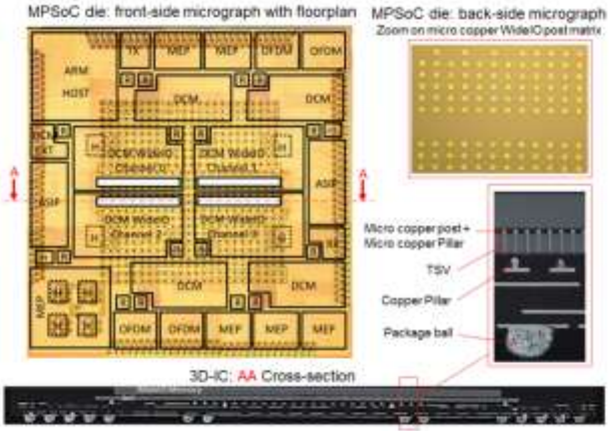


Figure 2 : 3D Integrated-Circuit microphotographs

As shown in figure 3, the silicon measured 208 MHz operating frequency at minimum voltage value (Vmin) with a high activity 13N MBIST test pattern proves the capability and robustness of our technology and design solutions as predicted by our simulations.

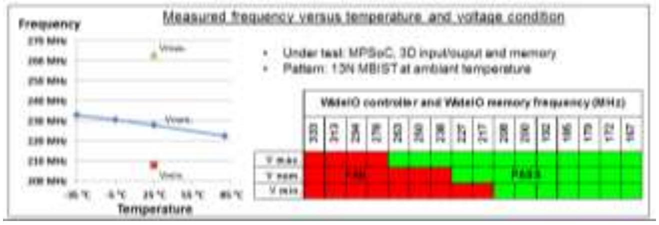


Figure 3 : Frequency characterization on silicon

During the peak activity section of the BIST pattern, we monitored the power consumption of the 3D I/O link. In such condition, we measured a power efficiency of 0.9pJ/bit which represents an improvement of a factor 4 with state-of-the-art LPDDR3 off-package I/O link.
3D stacking is a unique opportunity enabling memory-interconnect evolution to higher bandwidth with improved power efficiency and paves the way for innovative 3D stacked distributed caches [2].

Related Publications :
[1] D. Dutoit, et al. "A 0.9 pJ/bit, 12.8 GByte/s WideIO Memory Interface in a 3D-IC NoC-based MPSoC", VLSI Circuits (VLSIC), 2013 Symposium on , pp.C22,C23, 11-13 June 2013.
[2] Clermidy, et al. "3D stacking for multi-core architectures: From WIDEIO to distributed caches'' Proceedings of the, 2013 IEEE International Symposium on Circuits and Systems, ISCAS 2013, 19 May 2013 through 23 May 2013, Beijing', 537-540

Fast and Accurate System-Level Thermal Modeling: Application to a Memory-on-Logic 3D Circuit

Research topics: Thermal Analysis; 3D Circuits; Architecture Exploration

C. Santos, P. Vivet, D. Dutoit, P. Garrault (DOCEA), N. Peltier (DOCEA), R. Reis (UFRGS)

ABSTRACT: In 3D architectures power densities are larger and thermal dissipation is reduced with thin circuit substrates including TSVs. Thermal behavior must be evaluated and mitigated as early as possible in the design flow. We present a thermal modeling methodology, based on DOCEA Power ATM tool, including homogenization techniques to take into account fine grain structures (bumps, etc) while reducing model complexity. The methodology has been applied to a Memory-on-Logic 3D architecture, and validated against silicon results. Fast model exploration is possible to validate thermal hotspots.

Modern electronic systems require more and more computing power which is achieved by technology scaling and architectural evolution such as MPSoC architectures. However the technology scaling comes with a densification of the power consumption, resulting in increased thermal dissipation. This is even more the case with 3D architectures where stacking multiple dies further increase power density within the 3D stack. Current Many-Core architectures bring also issues of their own in the form of dynamic behavior in the power dissipation profiles, limiting systems' thermal predictability. A thermal mitigation solution must be developed in order to control both power and thermal dynamic profiles. As a consequence, chip's thermal behavior must be evaluated and mitigated as early as possible in the design flow. It is not acceptable to validate power and thermal budget at circuit sign off phase.

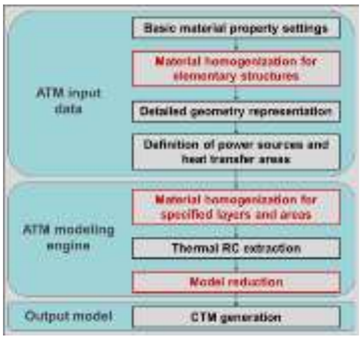

ATM input data
Basic material property settings
Material homogenization for elementary structures
Detailed geometry representation
Definition of power sources and heat transfer areas
ATM modeling engine
Material homogenization for specified layers and areas
Thermal RC extraction
Model reduction
Output model
CTM generation

Figure 1 : Proposed Thermal Modeling Methodology

In this work, we present an efficient thermal modeling methodology [1], based on DOCEA Power ATM tool, which allows taking into account both microscopic structures such as micro-bumps and TSVs, and macroscopic structures such as mother board power consumption or circuit package. The ATM thermal model (Figure 1) is based on a description of physical dimensions and thermal properties of all system elements (die floorplan, 3D stack thickness, micro-bumps and TSV locations, die package, die socket, and board structure). Associated to the thermal model is attached the corresponding power sources, with user's power scenarios. Low level accuracy of fine grain structures is achieved with homogenization techniques to accurately take into account all fine grain structures (bumps, TSVs, etc) while reducing thermal model complexity. The achieved compact thermal model (figure 3) can be simulated both in static and dynamic modes for system level exploration and optimization of thermal effects at various levels: die floorplanning, package selection, power profiles, etc.

The proposed methodology has been applied to WIOMING (figure 2) [2], 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 matrixes to connect to a WideIO compatible DRAM memory. The circuit is instrumented with thermal sensors and resistive heaters to emulate hot spot power dissipation, thus offering a full thermal characterization environment using embedded software on the real 3D circuit.

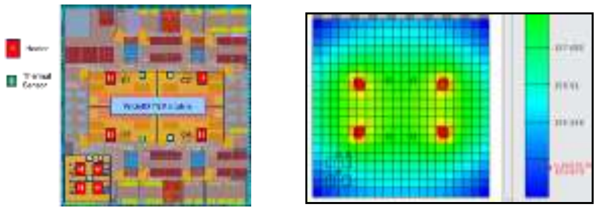


Figure 2 : WIOMING Circuit Floorplan, and Thermal Map with Heaters

Thermal static analysis (figure 3) shows very good accuracy for Hotspot behavior (less than 5% error for thermal measure versus thermal simulation). The thermal model is able to take into account fine grain structures (bumps, TSV) and system level constraints (board power consumption).

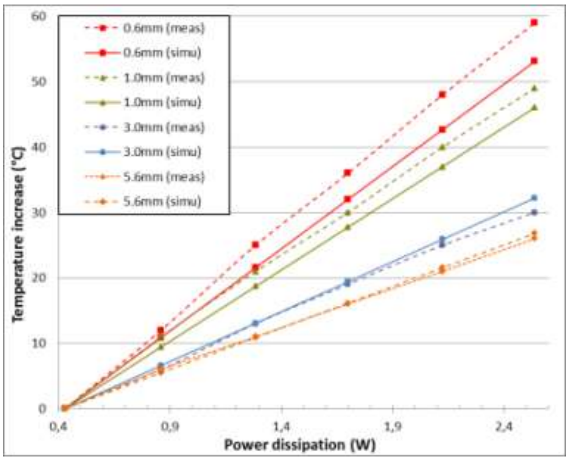


Figure 3 : Thermal Static Analysis: Simulation versus Silicon

Related Publications
[1] C. Santos, P. Vivet, D. Dutoit, P. Garrault, N. Peltier, R. Reis, " System-Level Thermal Modeling for 3D Circuits: Characterization with a 65nm Memory-on-Logic Circuit", IEEE 3D Integrated Circuit Conference, 3DIC'13, Oct 2013.
[2] D. Dutoit et all., "A 0.9 pJ/bit, 12.8 GByte/s wideIO memory interface in a 3D-IC NoC-based MPSoC," VLSI Symposium, Kyoto, Japan, June 2013.

Design For Test Architecture of TSV-based 3D Integrated Circuits

Research topics: 3D IC DFT architecture, TSV testing, BIST, IEEE 1687 (IJTAG)

Y. Fkih(Leti & LIRMM), P. Vivet, B. Rouzeyre (LIRMM), J. Schloeffel (Mentor Graphics)

ABSTRACT: 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 test issues, especially for TSVs. We propose in this work a novel Built-In-Self-Test (BIST) architecture for pre-bond testing of TSVs in 3D stacked integrated circuits, and a 3D DFT architecture based on IEEE 1687 (IJTAG) standard in order to test all components of the 3D system.

The proposed 3D DFT architecture addresses a first issue, the test of TSVs in Pre-bond phase, where TSVs are accessible from only one side. For the pre-bond test of TSVs, the test architecture is based on ring oscillators, allowing to measure TSV capacitance variations. In fact a variation on the oscillation frequency means a variation on the TSV capacitance. The detection of the frequency variation is done by performing a comparison between frequencies of all ring oscillators in the circuit according to a test strategy.

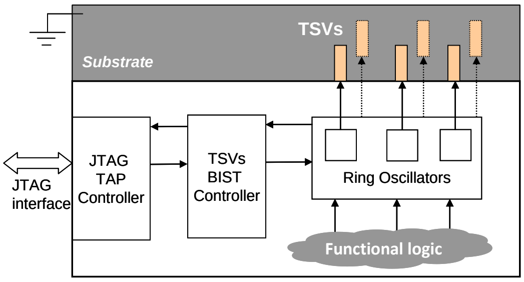


Figure 1: TSV Test principle using BIST and ring oscillators

Figure 1 shows the test architecture. TSVs are integrated in the silicon substrate, which is biased to the ground, and TSVs are not accessible from the external side. In pre-bond, before 3D stacking, only one side of the TSV can be accessed from the logic side which is used to perform the test. The test architecture is composed of two parts : i) the ring oscillators charging and discharging TSVs, frequencies of which depend on TSV capacitances ; and ii) the BIST controller that permits the control of the test, and the generation of test signatures that can be easily captured and shifted out through a standard JTAG interface. The design has been implemented in CMOS 65nm and uses only standard-cells.

In addition to the pre-bond test of TSVs, a 3D test architecture is proposed to test all other components of the 3D system. The 3D test requirements are the following: be compatible with existing standards, allow pre-bond and post-bond test, offer enhanced test concurrency within the 3D stack, and the possibility of test pattern retargeting from 2D to 3D. To respond to these test requirements, the 3D DFT architecture is based on the IEEE 1687 (IJTAG) standard as shown in figure 2. Each die integrates a TAP controller 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 (TSVs) for post-bond level.

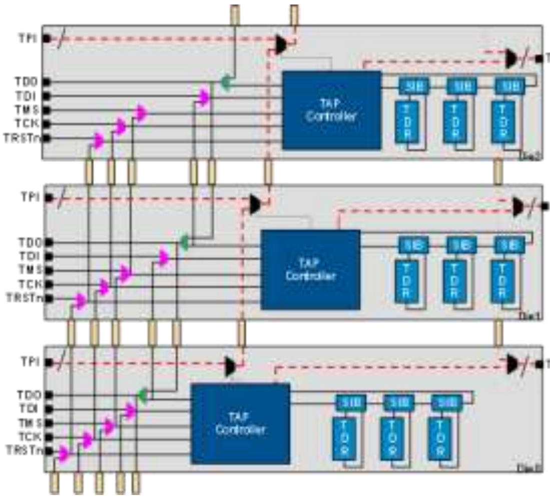


Figure 2. IJTAG based 3D DFT architecture

The IEEE P1687 (IJTAG) is an extension of the existing IEEE 1149.1 JTAG standard, but offering two main advantages : i) easy test pattern retargeting from 2D (for 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.

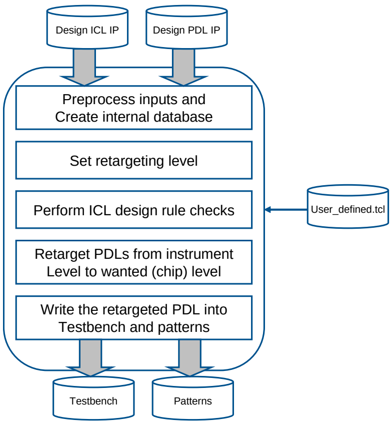

Design ICL IP
Design PDL IP
Preprocess inputs and Create internal database
Set retargeting level
Perform ICL design rule checks
User_defined.tcl
Retarget PDLs from instrument Level to wanted (chip) level
Write the retargeted PDL into Testbench and patterns
Testbench
Patterns

Figure 3. Tessent IJTAG flow

For 3D DFT insertion and ATPG, Tessent IJTAG tools from Mentor Graphics is used. The flow of the tool is shown in figure 3, where input files are ICL (for description of 3D DFT structure) and PDL (for test procedures). The use of high level languages allows to reduce test development time and increases the reusability of test patterns from 2D test to 3D test.

Related Publications:
[1] Y.FKIH, P.VIVET, B.ROUZEYRE, ML.FLOTTES, G.DINATALE "A 3D IC BIST for pre-bond test of TSVs using ring oscillators" IEEE International New Circuits and Systems Conference (NEWCAS), 2013
[2] Y.FKIH, P.VIVET, B.ROUZEYRE, ML.FLOTTES, G.DINATALE "A JTAG based 3D DfT architecture using automatic die detection", IEEE PRIME, pp 341-344, 2013 (bronze leaf award)

Programming for Future 3D Manycore Architectures: The PRO3D & SMECY Approach

Research topics : 3D ICs, Manycores, Thermal integrity, Software

C. Fabre, S. Bensalem(UJF), E. Flamand(ST), L. Benini(Unibo), L. Thiele(ETHZ), D. Atienza(EPFL)

PRO3D (FP7, http://pro3d.eu) tackled two important 3D technologies, that are through silicon via (TSV) and liquid cooling, and investigated their consequences on stacked architectures and entire software development. SMECY (Artemis) started on the grounds that performance could only be mastered using a design approach that optimizes interaction between SoC design and Embedded Software approaches. As a key result, a software design flow based on the rigorous assembly of software components and monitoring of the thermal integrity of the 3D stack has been developed by us and other partners.

With the ever increasing demand for higher data rates and performance as well as multi-functional capabilities in circuits, vertical integration of IC dies using through-silicon vias is envisioned to be one of the most viable solutions for the development of new generation of electronic products. 3D integration of multi-core processors offers massive bandwidth improvements while reducing the effective chip footprint. However 3D integration introduces several challenges, mostly related to the following factors: (1) increasing amount of logic that can be placed onto a single 3D IC; (2) Related thermal dissipation problem; (3) A necessary shift in programming models towards more parallelism. A software design flow [1, 2, 3] provided development tools and runtime to address these concerns heads front - Fig. 1.

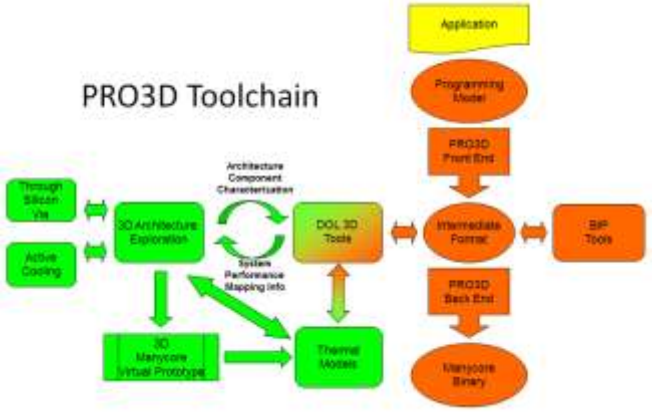


Figure 1 : The PRO3D Toolchain.

The new issues introduced by 3D stacking are mostly related to thermal issues. These issues have two main origins: (1) Thermal cross-coupling of execution units. The relative position of processing units as a whole, or computing units from therein (operators, instructions decoders, register files, caches, etc.) and memory defines how heat from one element impacts another one. If two processors are too close to each other, we may have to offload both of them in situations where a single one could have run without harm. So not only the topoplogy of the manycore will have to be known from the compilation flow and the runtime, but also the geometry and thermal characteristics of the hardware, (2) Different time scale for thermal propagation and computation forecast. Manycore architectures are in the GHz range, while the evolution of the temperature is in the Hz range. This means several orders of magnitude between the cause of heating (computations) and heating itself [6]. This gap in dynamic magnitude is reinforced by the fact that even at constant frequency, energy consumption increases with temperature. All this makes it difficult to reverse temperature variations. Any decision related to thermal management will probably have to use predictive thermal models.

Impact on Runtimes & Programming Models

The consequence of 3D stacking on the runtime and the programming model are twofold: (a) The fading of pure static compilation: due to the huge gap of time scale between computation and thermal effects, it seems very difficult, if doable at all, to build fully-static compilation schemes where the compiler will decide of the mapping offline, before execution, once and for all.

At least to ensure the platform's thermal integrity, some level of responsibility w.r.t. mapping must be left to the runtime To ensure this integrity the runtime will have to deal with tasks scheduling and resource allocation while taking into account not only the architecture's topology and the computation load, but also the actual geometry and thermal characteristics of the material involved in the architecture. This will require programming models that can provide enough flexibility at execution whereas essential properties can be guaranteed at compile-time [11]. (b) The fading of von Neumann as a programming model: as for programming models, we should move away from von Neumann (only as programming model, not as computing Architecture) and consider other kinds of programming models naturally parallel, like process network and message passing already discussed. Even these parallel programming models must be checked to be amendable to analyses that can predict the amount of computing load, if not to an absolute time reference, at least towards a moving horizon. This is necessary to provide computation forecasts to a runtime scheduler that can efficiently use the stacked architecture while preserving its thermal integrity.

Related Publications:
[1]C. Fabre et al., "PRO3D, 'Programming for Future 3D Manycore Architectures: Project Interim Status', 10th International Symposium FMCO 2011. State-of-the-Art Survey. Bernhard Beckert, Ferruccio Damiani, Frank de Boer, and Marcello Bonsangue, editors. volume 7542 of LNCS. Springer, 2013.
[2]J. Mottin et al., 'Compilation tool chains and Intermediate Representations', chapter of Torquati, M.; Bertels, K.; Karlsson, S. & Pacull, F., ed. (2013), "Smart Multicore Embedded Systems", Springer, The Netherland, pp. 21-32.
[3] "The STHORM platform", Mottin, J.; Cartron, M. & Urlini, chapter of Torquati, M.& al. (2013), pp 35-43.

An Efficient and Flexible Hardware Support for Accelerating Synchronization Operations on the STHORM Many-Core Architecture

Research topics: Synchronization, Many-core, HW acceleration, STHORM

F. Thabet, Y. Lhuillier, C. Andriamisaina, J-M. Philippe and R. David

ABSTRACT: The current trend in embedded computing consists in increasing the number of processing cores on a chip. Synchronization handling on this architecture was critical since speed-ups of parallel implementations of embedded applications strongly depend on the ability to exploit parallelism. This paper presents the HardWare Synchronizer (HWS), a flexible hardware accelerator for synchronization operations, developed for the STMicroelectronics/CEA STHORM architecture. Experiments on a multi-core test chip showed that the HWS has less than 1% area overhead while reducing synchronization latencies (up to 2.8 times) and contentions.

Synchronization handling is critical in embedded many-core platforms like STHORM [3] since synchronization events are frequent, especially when using fine-grain parallelism (small job scheduling under time and power constraints).

The proposed HardWare Synchronizer (HWS) [1, 2] was introduced to address challenges such as:
- Acceleration of common synchronization constructs,
- Flexibility and composability to adapt to evolving execution models,
- Ease of integration in most systems and area efficiency,
- reducing bandwidth, power consumption and polling,
- Ability to be replicated in many-core systems (scalability),
- Providing support for atomic operations in absence of atomic support in the memory sub-system

The HWS module is an area-efficient/scalable hardware accelerator for common synchronization primitives. It is designed as a shared peripheral (a set of registers) to be seamlessly integrated in all architectures using load/store operations.

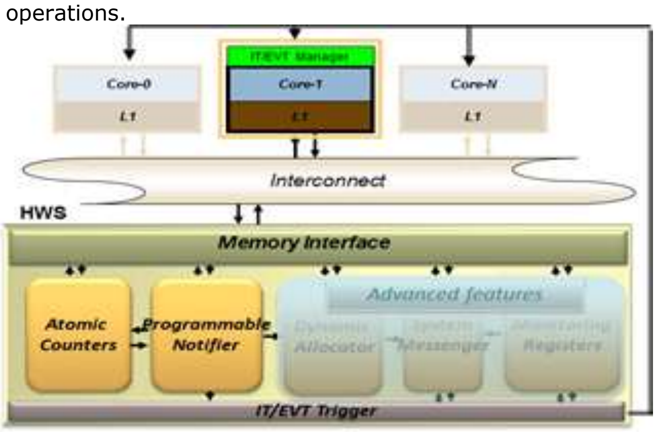


Figure 1 : HWS architecture overview.

The HWS is modular: dedicated blocks implement functions supporting synchronization and communication primitives:

- Atomic Counters (AC): atomic registers allowing to implement a wide range of atomic operations / synchronization mechanisms such as mutexes, semaphores and barriers.

- Programmable Notifier: enables computing cores to schedule event notifications according to specific ACs values so as to avoid software polling.

- Interrupt/Event Trigger: generates interrupt/event to cores.
- Advanced Features: hardware-assisted mechanisms for handling dynamic task-scheduling, resource allocation acceleration and thread migration (not in the scope of the presented work).

The HWS was chosen to be the infrastructure for synchronization handling in the STHORM platform and in the CEA internal Locomotiv [2] test chip. The Figure 2 and tables I show some area and power characterization results performed under ST-CMOS 32nm lib, at 500 Mhz frequency for PE (Processing Element).

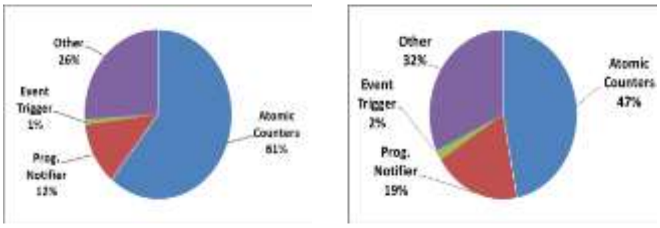
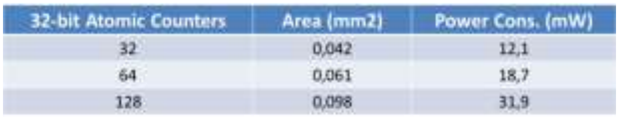


Figure 2 : Contribution of different HWS parts to the global area in a 32nm technology: (a) Locomotiv Configuration – 4 PEs and 64 32-bit ACs, (b) STHORM Typical Configuration - 16 PEs and 128 32-bit ACs



32-bit Atomic Counters	Area (mm2)	Power Cons. (mW)
32	0,042	12,1
64	0,061	18,7
128	0,098	31,9

Table I : Area and power consumption of the HWS with respect to the number of ACs

Runtime software acceleration: experiments on the Locomotiv test chip showed that the HWS has less than 1% area overhead while reducing synchronization latencies (up to 2.8 times) and contentions.

Related Publications:
[1] F. Thabet, Y. Lhuillier, C. Andriamisaina, J-M Philippe and R. David, " An Efficient and Flexible Hardware Support for Accelerating Synchronization Operations on the STHORM Many-Core Architecture," DATE 2013, pp.531-534, March 2013.
[2] E. Beigne, I. Miro-Panades, Y. Thonnart, L. Alacoque, P. Vivet, S. Lesecq, D. Puschini, F. Thabet, B. Tain, K. Benchehida, S. Engels, R. Wilson, D. Fuin. "A Fine Grain Variation-Aware Dynamic Vdd-Hopping AVFS Architecture on a 32nm GALS MPSoC", ESSCIRC-2013 conference, San Jose, CA USA, Sep 14-17, 2014.
[3] L. Benini, E. Flamand, D. Fuin, and D. Melpignano, “P2012: Building an ecosystem for a scalable, modular and high-efficiency embedded computing accelerator,” in Design, Automation & Test in Europe Conference & Exhibition (DATE12), 2012, pp. 983–987. [2]….

Worst case latency calculation in Data-Flow programs for clusters of the MPPA Manycore chip

Research topics: Worst-Case latency, Communication overhead, Shared memory, Many-core, Data-Flow.

A. Dkhil, S. Louise, C. Rochange (IRIT)

ABSTRACT: For real-time applications, dealing with multi-core or many-core systems is known to be difficult, especially with regards to worst-case execution times (WCET). The additional difficulty concerning WCET is due to the sharing of resources. In this paper, we show how to calculate the worst-case communication overhead for accessing shared memory clusters of MPPA chip from Kalray (a many-core processor with 256 cores that appeared last year). This is done in order to compute end-to-end latency for stream programs.

Stream programming and data-flow concepts are gaining momentum for embedded high performance programming of many-core systems. The original motivation for research into dataflow was the exploitation of massive parallelism, mainly in the DSP domain. What is required to know to calculate worst case latencies in addition to stand-alone WCETs of individual tasks is the communication overhead induced by the interference when several processors want to access simultaneously to the same bank of shared memory. This is a first evaluation of a theoretical development for CycloStatic DataFlow graphs, the mathematical base of the ∑C language [1].

The MPPA many-core has 256 cores available for user computation. They are organized as 16 clusters of 16 processors whose clusters can communicate with the others thank to a Network on Chip. Within a cluster, the main medium of communication is a banked shared memory (Fig. 1).

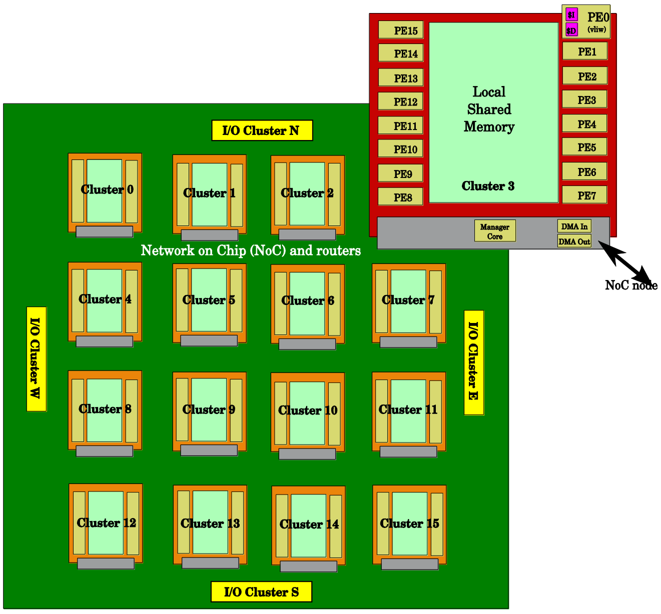


Figure 1 : A simplified view of Kalray's MPPA architecture

The banked memory is implemented in a multi-bus approach. Each memory-bank has its private controller which manages the requests sent from each processor in the cluster using a FIFO (first-in, first-out) queuing strategy. The minimum time needed to satisfy a request t_0 is constant. The access time is $t = t_0 + (\alpha - 1)t_c$ where $(\alpha - 1)t_c$ models the arbitration cost of the memory controller. This cost is not constant: it models the order of processor requests in the FIFO of the controller. α is the number of cores in a cluster. t_c is the cycle time of memory.

A dataflow graph comprises a number of filters (processes) performing computations and one or several channels in and out for communication between filters. We have implemented an algorithm that computes the end-to-end latency in such a dataflow graph. For that we based our study on a static scheduling of the set of filters which is always possible for memory bounded CSDF applications. It takes into account data dependencies and memory access and communications between filters.

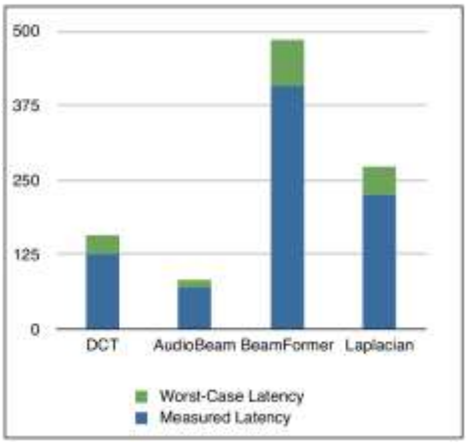


Figure 2: Comparison between measured and worst-case latency predicted by our model, on several usual dataflow applications

Not all the memory accesses for a given core will have the maximum overhead. As a consequence, for the relevant cases, the worst-case overhead is between 17% and 23% of the latency derived from simulation results (Figure 2), which is a tight estimation compared to the state of the art [2].

Related Publications:

[1] T. Goubier, R.Sirdey, S. Louise, and V. David. "∑C: A programming model and language for embedded manycores", ICA3PP 2011.
[2] A. Dkhil, S. Louise, C. Rochange, " Worst case latency calculation in Data-Flow programs for clusters of the MPPA Manycore chip", 7 th Junior Researcher Workshop on Real-Time Computing JRWRTC 2013, co-located with RTCSA 2013.

A formal evaluation of mean-time access latencies for interleaved on-chip shared banked-memory in manycores

Research topics : Manycore architecture, memory-access delay, mathematical model

Stéphane LOUISE

ABSTRACT: Manycore architectures are the future of embedded system, nonetheless, mastering their huge processing power while ensuring safety-critical properties brings new challenges. One of them is to characterize access-time delays to on-chip memory when several cores try to access the same memory banks concurrently. As on-chip memories of embedded manycores are usually single-port for the sake of power consumption, the resulting serialization of memory accesses causes latencies. We have built a theoretical model of these latencies for multibanked memories which would be useful for establishing real-time and schedulability properties of manycores.

Several manycore chip for embedded systems appeared in the last couple of years. Two notable chips are the STHORM architecture from ST Microelectronics and MPPA from Kalray. Both have some similarities: clustered architecture which can communicate via a Network on Chip; Cluster is composed of 16 processors (or Processing Elements, PE) and a banked shared memory. Any PE can access any bank in the cluster, and the memory addressing scheme in interleaved so that no bank would become a hot spot: all memory accesses are spread naturally on all the banks of the cluster.

Real-time, scheduling, and performance analyses require that several properties are known. One of the important properties for characterizing such a system is the access time to memory. For embedded systems like STHORM or MPPA, memory banks are single-port memories so, at most, only one access is granted each cycle to a given bank. To ensure a fair and deterministically time-bounded access to any bank, a round-robin arbitration is used. And as clocks even within a cluster are not absolutely synchronized (locally synchronous, globally asynchronous), we can modelize delays of first access time to a bank as a roulette game outcome, i.e. a random variable noted D.

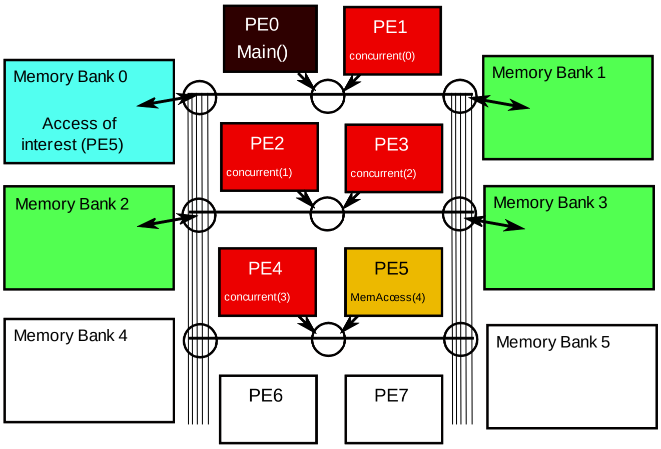


Figure 1: Partial view of a cluster with 8 cores and 6 memory banks. The interest thread is on PE 5 experiments concurrent accesses from PE 1, 2 and 3.

The diagram on figure 1 shows the case when n PEs (n=5 for Figure 1) try to access up to m memory banks concurrently (m=4 on Figure 1). By doing a probabilistic approach we found a general expression of the mean delay $\overline{D_{n,m}}$ for any values of n and m. We also showed a tight approximate upper bound of this delay as an analytical expression:

$$\overline{D_{n,m}} \approx \frac{\alpha}{2}\left(\left(1+\frac{1}{m}\right)^{n-1}-1\right)$$

Where α account for memory access rate of PEs. In the case of the MPPA architecture, n=16 and m=16, which gives as mean delay $\overline{D} = 0.7$ cycles. We also made several evaluations for different utilizations of the cluster, and several configurations of the architecture, in order to find out when this memory architecture behaves nicely and when it does not. This result is summarized in Figure 2.

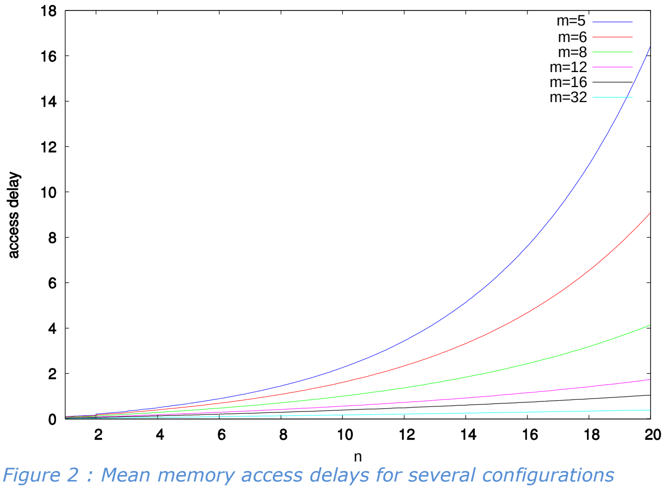


Figure 2 : Mean memory access delays for several configurations

As can be seen, such memory architecture behaves well (mean delays less than 1 cycle) as soon as the number of banks m is at least the number of cores n ($n \leq m$), and behaves reasonably well if n<2m. Future works will validate these results experimentally and also take the standard deviation into account.

Related Publications:
[1] S. Louise, "A formal evaluation of mean-time access latencies for interleaved on-chip shared banked-memory in manycores" IEEE International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC-13), Tokyo 2013.

ARTM: A Lightweight Fork-Join Framework for Many-core Embedded Systems

Research topics : low overhead scheduling, many-core architecture

M. Ojail, R. David, Y. Lhuillier, A. Guerre

Embedded architectures are moving to multi-core and many-core concepts in order to sustain ever growing computing requirements within complexity and power budgets. Programming many-core architectures not only needs parallel programming skills, but also efficient exploitation of fine grain parallelism at both architecture and runtime levels. Scheduler reactivity is however increasingly important as tasks granularity is reduced, in order to keep the overhead of the scheduling to a minimum. We present a lightweight fork-join framework for scheduling fine grain parallel tasks on embedded many-core.

In order to address the constant demand for increasing performance within the complexity and power budgets of embedded systems, current computing architectures are becoming massively parallel systems-on-chip. Many-core architectures intend to move away from the trend of increasing performance through complex micro-architectures that support instruction-level parallelism (ILP), and embracing designs with multiple, simple cores on a chip to exploit task-level parallelism (TLP) and data-level parallelism (DLP). This change in computer architectures enables processors to be clocked at a lower frequency and to consume less power, while still getting better overall performance.

Fine grain parallelism is required for certain applications since it eases the work of the developer being a form of parallelism which is naturally present in applications and does not require heavy algorithm rewriting. Since coarse grain threads can limit opportunities for exploiting parallelism for those applications, an efficient way to use the resources of multi-core platforms, is by exploiting the inherent fine grain parallelism. Meanwhile scheduler reactivity is becoming increasingly important as tasks granularity is reduced, in order to keep the overhead of the scheduling to a minimum. Thus, solutions based on the creation of new threads for more parallelism are not feasible since they induce too large time overheads in case of fine-grain tasks.

However, the fork-join model of parallelism, already used for coarse grain threads, remains among the simplest and most effective design techniques for obtaining good parallel performance. We thus propose in [1] a lightweight, Asynchronous Reactive Tasks Management (ARTM) framework, based on the fork-join model to exploit fine grain parallelism at the lowest possible cost.

Figure 1 shows an example of a task graph based on the ARTM framework. In this figure, the master core forks 3 tasks T0 to T2, then duplicates 2 times the task T3 before going into slave mode. Tasks T3 will execute M times before joining, while tasks T0, T1 and T2 will execute N times before joining. Task T0 will only execute every other time and forks tasks T4 and T5 each time it is executed.

In Figure 2, forking and joining tasks with the ARTM framework is schematically explained via an example. It shows a snapshot of the main table for the task graph depicted in Figure 1 where only T0 has begun execution.

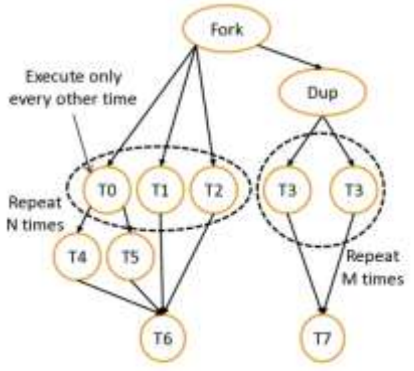
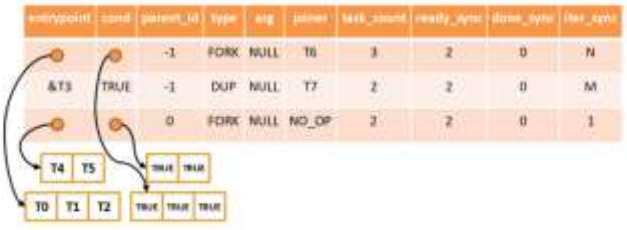


Figure 1 : Task graph example based on the ARTM framework

Three concurrent fork/dup operations are in the system. Since only one element of the main table is needed per fork/dup operation, scheduling weight is independent of the number of forked or duplicated tasks.



entrypoint	cond	parent_id	type	arg	joiner	task_count	ready_sync	done_sync	iter_sync
		-1	FORK	NULL	T6	3	2	0	N
&T3	TRUE	-1	DUP	NULL	T7	2	2	0	M
		0	FORK	NULL	NO_OP	2	2	0	1

T4
T5
TRUE
TRUE
T0
T1
T2
TRUE
TRUE
TRUE

Figure 2 : Snapshot of the main ARTM table for the task graph in Figure 1

As a first experiment, performance estimation of the code is done on a cycle accurate ISS of one STxP70-V4 processor of STMicroelectronics. We test the scheduling loop in two cases: scheduling tasks corresponding to the same fork/dup operation and scheduling tasks corresponding to different forks. In the first case, the overhead of the scheduling loop is 26 cycles while in the second case, this overhead increases to 29 cycles. Concerning the fork/dup operations, the overhead is 35 cycles for the top operations (with a parent ID of -1) and 43 for the other operations. This difference of 8 cycles is due to an extra atomic post-increment operation at the parent level. As for the tasks joining, the simulations on the ISS show an overhead of 40 cycles per join operation. More experimentations and results are available in [1].

Related Publications:
[1] M. Ojail, R. David, Y. Lhuillier, A. Guerre, "ARTM: A Lightweight Fork-Join Framework for Many-core Embedded Systems", Design, Automation & Test in Europe Conference & Exhibition (DATE), 2013.

A GRASP for placement and routing of dataflow process networks on manycore architectures

Research topics: dataflow compilation, placement and routing, GRASP heuristic

O. Stan, R. Sirdey, J. Carlier (UTC), D. Nace (UTC)

ABSTRACT: In this study, we propose a GRASP heuristic for solving the joint problem of placement and routing of process networks from the field of compilation for embedded manycore architectures. The method we propose consists in assigning applications expressed as dataflow process networks on homogeneous manycore architectures by taking into account the routing maximal capacity of the arcs for the underlying Network-On-Chip. Our experiments, also validated on a real embedded application, illustrate the algorithm ability to efficiently obtain good quality routable assignments, within an acceptable computational time.

In order to efficiently exploit the parallelism and to take full advantage of the computing power the manycore systems may provide, their design requires new programming and execution paradigms as well as innovative compilation technologies. Dataflow paradigm seems to be a good candidate for programming manycore applications.

In dataflow models, an application is described as a static instantiation graph of concurrent tasks interacting through unidirectional FIFO channels. The compilation process of a dataflow application for a manycore architecture is becoming rather complex and requires solving a number of difficult and large-size optimization problems in order to efficiently allocate and exploit the inter-related resources.

The optimization problem we consider in [1], related to the resource allocation pass of compilation, consists in the joint placement and routing of Dataflow Process Networks (DPN) on a homogeneous clusterized manycore architecture in which the cores are organized as clusters communicating through an asynchronous Network-On-Chip (NoC).

Figure 1 gives an overview of the main components of a dataflow graph for a motion tracking application, used for validation, as well as of the target clusterized architecture, with a 2D 4x4 torus topology.

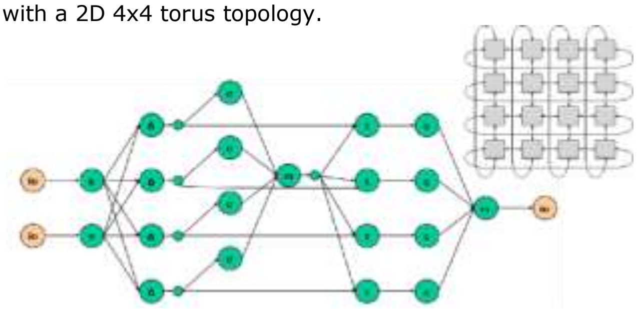


Figure 1: Motion target application to be placed on a 2D 4X4 Torus.

Even if the two sub-problems of tasks mapping and routing have already been addressed in the literature, the novelty of our approach consists in treating together task mapping and routing, and thus, taking into account the routing when placing the networks of processes.

We study the static mapping of tasks from a DPN onto a network of clusters, such as the total bandwidth used by the application is minimal and there exists a routing path between tasks situated on different clusters. we are interested in finding an admissible assignment of tasks to clusters minimizing the sum of inter-tasks communications. In the context of our work, an admissible assignment is a mapping of tasks to clusters which satisfies the capacity constraints for each cluster and each resource and, furthermore, it assures that there exists a feasible routing between every two communicating tasks, which route respects the maximal capacity of the links of the network.

Since the tasks mapping is equivalent to the NP-hard Node Capacitated Graph Partitioning problem and the unsplittable flow problem can be restricted to the Directed Edge Disjoint Paths problem, also NP-hard, the joint problem is straightforwardly NP-hard in the strong sense.

Taking into account the problem complexity and the size of the instances we have to deal with, we turned our attention to approximate algorithms and in particular to the GRASP (Greedy Randomized Adaptative Search Procedure) metaheuristic, which seems a more suited choice to tackle this problem especially for the beginning of the development cycle of an application.

GRASP is a multi-start metaheuristic, with each iteration involving two phases: construction and local search. The main idea of our constructive part is to verify at each step of the mapping, that the flows between the assigned tasks can be routed by making use of the previous computed flows. At each step of the mapping, the computation of new routing paths is realized through a single source shortest-path algorithm on a reduced graph obtained from the original NoC and whose arcs are weighted with a residual capacity. Each time, a decision is chosen randomly between a list of k best decisions constructed in a greedy fashion and consisting of the admissible assignments of tasks to clusters and of the fusions of two clusters with the highest relative affinity. Afterwards, the quality of the constructed solution is improved through a local search procedure, by generating a new solution through the interchange of pairs of tasks assigned to different clusters.

For testing our algorithm, we used several sets of test problems: several grids to be placed on square grids, a modified version of Johnson instances and the motion target application. The results show that our algorithm is able to find better solutions than the sequential algorithm (placement followed by routing) and it is able to propose routable placements event when the capacity on the arcs on the networks is limited.

Related Publications:
[1] O. Stan, R. Sirdey, J. Carlier and D. Nace, "A GRASP for placement and routing of dataflow process networks on manycore architectures", Proceedings of the 8th IEEE International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 219-226, Compiègne, France, 2013.

SNet, a flexible, Scalable NETwork paradigm for manycore architectures

Research topics: Manycore architectures, Scalable network, Ant Colony Optimisation.

C. Azar, S. Chevobbe, Y. Lhuillier, J-P Diguet

ABSTRACT: A scalable communication paradigm for manycore architectures, called SNet (Scalable NETwork), is presented. It offers a wide range of flexibility by exploring the routing paths in a dynamic way, taking into consideration the network load. It is then followed by the data transmission phase through the chosen path.

One among the limits of the gigascale SoCs (System-On-Chip) will be the ability to efficiently interconnect pre-designed functional blocks and to accommodate their communication requirements in a scalable manner [2]. We propose to dedicate a set of Processing Elements (PE) to achieve routing tasks and a new PE co-processor, called DMC (Direct Management of Communications), to handle data transfers. With simple software libraries uploaded in PE memories, the user defines a specific topology, which is based on existing communication links between PEs. Then the routing is dynamic and implemented in a distributed manner in order to create paths from producer to consumer according to data dependencies. The whole concept is called Snet.

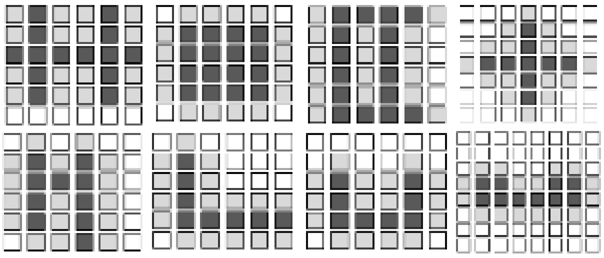


Figure 1: Routing networks tested on the manycore platform. Dark gray refer to routing nodes, light gray to computing nodes, and white to unassigned cores.

The routing strategy in SNet, implemented in the software library, is based on a distributed ACO (Ant Colony Optimization) algorithm in which dynamic paths exploration is achieved in parallel for all communicating tasks. More details on the platform and the ACO algorithm are found in [3].

We follow hereby an evaluation methodology developed in [4] to extract the characteristics of different routing topologies applied to SNet, by setting fixed simulation parameters to all tested topologies and modeling traffic with Poisson injection distribution. 16 computational nodes are mapped to the platform, as shown in Figure 1, with a varying number of routing nodes.

Each node of the platform injects 640 messages of n flits in the network, the flit width being 32 bits. In order to test the limits of the network saturation, n varies from 10 to 1000 flits, which is equivalent to 25 KB and 2.44 MB respectively. The simulations are conducted on our ISS simulator, which is based on an ISS array of RISC processors and model one instruction per cycle.

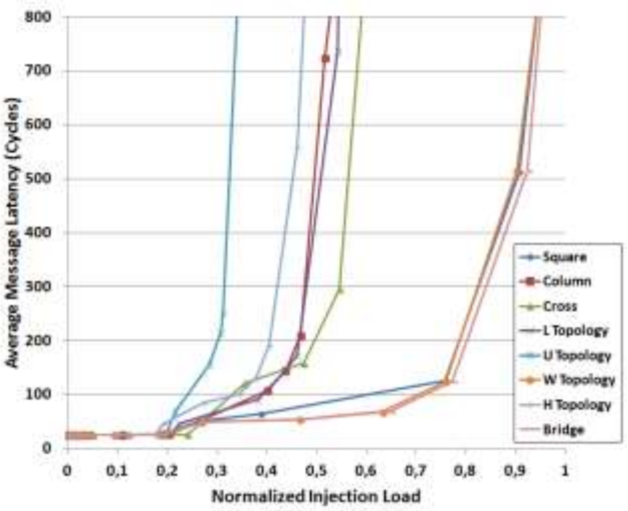


Figure 2: Latency variation following the Poisson injection distribution.

Measured latencies for the studied routing topologies under spatially uniform traffic distribution are plotted in Figure 2 for best mapping in which each two communicating tasks are placed as close as possible. In the actual implementation, the network presents a bandwidth of 9.6 Gbps at 300 MHz, considering that one transmission can be achieved simultaneously by the DMC module.

SNet efficiently implement a novel routing paradigm designed for scalability, flexibility, and ease of programmability.

It performs path exploration before data transmissions using the ACO distributed algorithm, without interrupting the application running. Results show that, though offering high flexibility and scalability, SNet manages to score high injection rate values reaching 0.9, and low message latencies. In the next project steps, time reduction of the ACO paths creation stage will be feasible by introducing specific hardware modules for ant agents propagation. Multitasking execution will increase computing efficiency and pave the way of dynamic migration.

Related Publications :
[1] Azar, C.; Chevobbe, S.; Lhuillier, Y. & Diguet, J.-P. , 'SNET, a flexible, scalable, network paradigm for manycore architectures', NOCS 2013
[2] Bertozzi, D. 'Xpipes: a network-on-chip architecture for gigascale systems-on-chip', Circuits and Systems Magazine, IEEE, no. 2, 2004
[3] Azar, C.; Chevobbe, S.; Lhuillier, Y. & Diguet, J.-P. , 'Dynamic routing strategy for embedded distributed architectures', ICECS 2011
[4] Pande; P 'Performance evaluation and design trade-offs for network-on-chip interconnect architectures' Computers, IEEE Transactions on, 2005

Hardened Asynchronous Networks on Chip

Research topics: Asynchronous circuits; Networks on chip; Data error correction

J. Hilgemberg Pontes, P. Vivet

ABSTRACT: Asynchronous circuits have shown better response under Single Event Effects than their synchronous equivalent. Contrary to what happens in synchronous circuits, delay variations induced by radiation usually have no impact on asynchronous Quasi-Delay Insensitive (QDI) circuits. However, bit flips may corrupt data transmissions and stall the circuit with no recovery solution. In order to overcome problems related to bit flips, the QDI asynchronous communication architecture is hardened at different design levels. As result, the communication robustness is leveraged providing high reliability and availability communication services.

In order to propose asynchronous Network on Chip for aggressive Multi-core architecture, tolerant to Single Event Effects, it is required to address hardening techniques at several design levels.

1) - Cell Level: A characterization method from standard cell to SoC level was proposed in this work. It enables cell swapping or library set generation for robust synthesis, and place and route.

2) - Logic Level: Circuit duplication and double check is performed in the most susceptible parts of the logic. The susceptible points are identified by the proposed simulation environment.

3) - Data Level: Data correction methods for m-of-n Delay Insensitive are proposed and implemented [1].

Figure 1 shows details of a communication architecture for a GALS SoC based on an asynchronous NoC that employs a data correction scheme. Two data corrections methods for m-of-n Delay Insensitive Encoding are proposed. The first method is based on parity calculation for m-of-n encoding. In this method, the unordered property of Delay Insensitive Encoding is used to detect errors. Combined with parity calculation for m-of-n encoding, the unordered property can be used to detect and correct errors in packet payloads. The second method (called Temporal Redundancy in Delay Insensitive Codes or TRDIC) uses the encoding space of m-of-n encoding to insert temporal redundancy to data. Both methods can be used together in a static way where the parity helps the TRDIC decoding or in a dynamic way where the data correction scheme is selected based on the error rate.

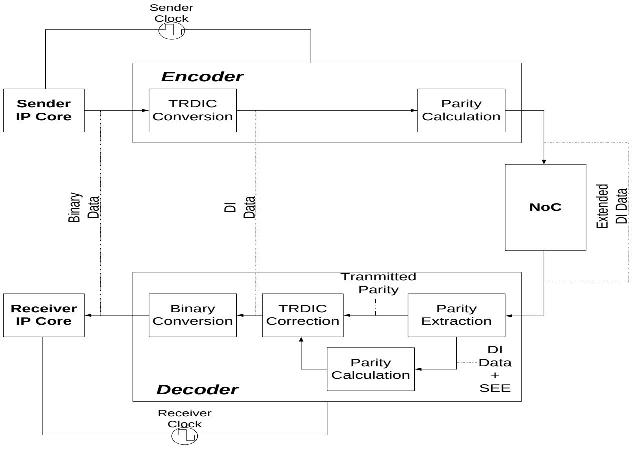


Figure 1 : Hardened architecture using Parity and TRDIC scheme

4) - Protocol level: to reduce the probability of stall in asynchronous handshake protocols the asynchronous buffer implementation was adapted to reduce the protocol timing windows most susceptible to soft errors [2].

5) - Packet Level: The NoC routing information and the Begin/End of packet indication are duplicated to ensure correct routing and packet integrity [2].

The flit type and the routing information are duplicated to increase the robustness of the packet control. Figure 2 shows the double check scheme adopted for the flit type and routing information. Thus, if an SEE occurs, double check (C-elements marked in Figure 2) is capable to filter it.

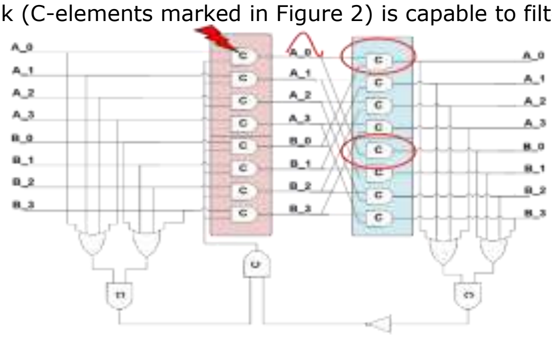


Figure 2 : Double Check redundancy within asynchronous pipeline

Figure 2 also shows the proposed Normal Open Pipeline Implementation (NOAP) [2], which corresponds to Hardening Implementation at Protocol Level. This implementation synchronizes the buffer input data and the acknowledge signal. As a result, the four phase protocol behavior changes reducing the probability of soft errors in C-elements that are part of the buffer structure.

The evaluation methodology was done by comparing the NoC with and without the proposed hardening techniques. Using 65nm technology, results show that the Hardened NoC reduces about x10 the probability of stall in the NoC when compared to a non-hardened asynchronous NoC [2]. The data correction scheme can deliver reliable data communication even in aggressive environments with low area/power cost [1]. Together, the proposed techniques can provide high level of robustness resulting in a high reliability and availability communication solution.

The continuation of this work includes evaluation of the proposed hardening techniques in FDSOI, since FDSOI technology also offers higher levels of robustness compared to bulk technologies.

Related Publications :.
[1] Pontes, J.; Vivet, P.; Calazans, N., "Parity Check for M-OF-N Delay Insensitive Codes," IEEE International on-Line Testing Symposium (IOLTS), May 2013
[2] Pontes, J.; Vivet, P.; Calazans, N., " H2A: A Hardened Asynchronous Network on Chip," Symposium on Integrated Circuits and Systems Design (SBCCI), September 2013

Simulation of BTI and HCI effects at high abstraction level

Research topics : BTI, HCI, Simulation, Timing degradation

O. Heron, C. Sandionigi, C. Bertolini, N. Ventroux and F. Marc (IMS-Univ., Bordeaux 1)

ABSTRACT: Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI) are responsible for timing degradation and possibly failures in Integrated Circuits (ICs). These ageing mechanisms are becoming more and more relevant due to die shrinking combined with non-ideal scaling of voltage. The evaluation of their effects early in the design flow becomes a must-have to ensure the expected time-to-market and IC's lifetime. This research aims at simulating BTI- and HCI-induced timing variations at high abstraction level. It lies on a bottom-up approach and, by considering processor microarchitectures, it analyzes the impact of instruction execution.

Ageing is becoming more and more a relevant problem in ICs due to the challenges brought by nano-scaling. This research studies the effects of two ageing mechanisms, BTI and HCI, on processors. When BTI and HCI are studied in literature, contributions usually consider low abstraction levels, where data are accurate but the time required to analyze a complex circuit is very high.

This research proposes a bottom-up approach that propagates knowledge about BTI and HCI effects from low abstraction levels to higher levels. At high level, it analyzes the impact of the Instruction Set Architecture (ISA) on the timing degradation of a processor. The final objective is to build a simulation framework for the exploration and verification of Multi-Processor Systems on Chip (MPSoC) under ageing conditions.

The approach computes the propagation delay of the processor's paths under ageing conditions. It is based on state-of-the-art models that calculate the delay of a gate. The models reveal how BTI and HCI depend on the processor's activity, i.e. Static Probability for BTI and Toggle Count for HCI. The flow to evaluate the paths' delay has been presented in [1] and is shown in Fig. 1. It analyzes the netlist of the processor to get the initial delay of the paths and it executes applications to annotate the activity. Finally, it applies the gate-level formula to obtain the ageing delay.

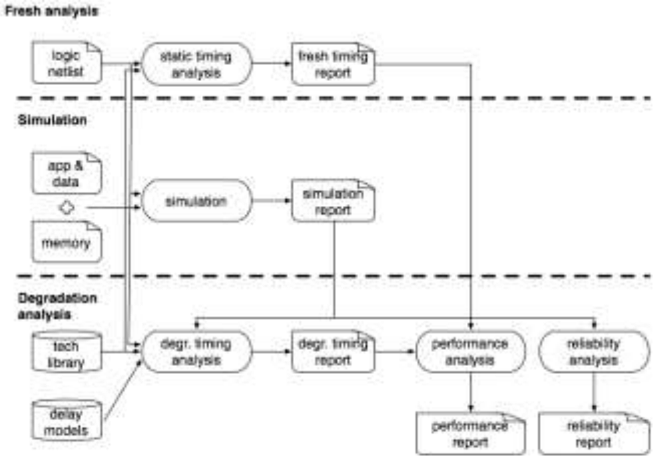


Figure 1 : Flow for the evaluation of ageing impact on processors

This flow is executed to characterize the processor's ISA and allow the analysis of timing degradation at high abstraction level. In [2], an approach to perform the analysis of HCI at functional level is described. It evaluates the impact of instruction execution, by exploiting a preliminary ISA characterization that annotates each path of the processor with a value of Toggle Count for each pair of instructions.

The proposed approach has been validated by considering the single processor core AntX and various typical embedded applications. The obtained experimental results are reported in [2]. The simulation at functional level allows to speed-up the analysis of timing degradation by a factor around 10 with respect to netlist simulation, as shown in Fig. 2. On the other hand, the functional level simulation underestimates the timing degradation. Despite the error, the approach allows to perform a comparative analysis of instructions and applications, hence identifying the ones causing the greatest degradation.

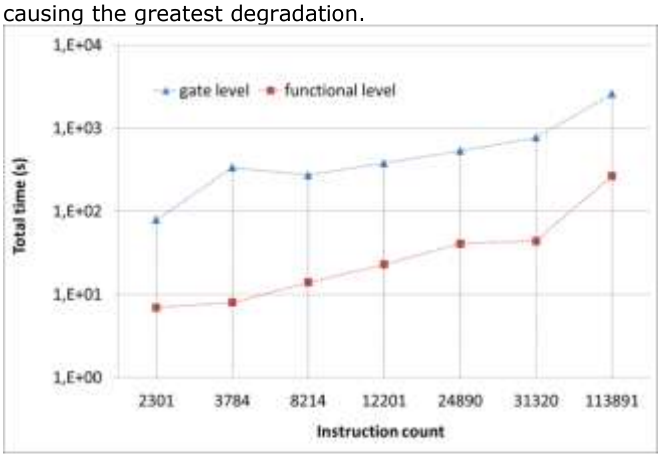


Figure 2 : Comparison of performance between netlist and functional simulation

Current work aims at refining the approach to get better accuracy in the evaluation of timing degradation and at analyzing the joint effect of reliability and temperature during the exploration phase of MPSoCs. Future work will consider the analysis of failures with stochastic nature, such as the joint effect of BTI and RTN (Random Telegraph Noise) in FD-SOI technology.

Related Publications:
[1] C. Sandionigi, O. Heron, C. Bertolini, R. David, "When processors get old: Evaluation of BTI and HCI effects on performance and reliability", IEEE International On-Line Testing Symposium (IOLTS), 2013, pp. 185-186
[2] O. Heron, C. Bertolini, C. Sandionigi, N. Ventroux, F. Marc, "On the simulation of HCI-induced variations of IC timings at high level", Journal of Electronic Testing: Theory and Applications (JETTA), 2013, Vol. 29, N. 2, pp. 127-141

Error-Correction Schemes with Erasure Information for Fast Memories

Research topics: MRAM; dielectric breakdown; erasure; ECC; shortened SEC; SEC-DED

S. Evain, V. Gherman

ABSTRACT: Two error correction schemes are proposed for binary memories that can be affected by erasures, i.e. errors with known location but unknown value. For example, erasures could be generated by the dielectric breakdown of a magnetic memory (MRAM) cell due to which the electrical resistance of the cell becomes smaller than the usual resistance values used to encode logic 0 and 1 values. Here, the erasure information is used to enable double-bit error correction (DEC) capability with the help of single error-correction double-error detection (SEC-DED) codes or shortened single error-correction (SEC) codes.

The robustness of memory subsystems can be effectively increased with the help of error-correction codes (ECCs). Unfortunately, ECCs are costly in terms of latency and storage overhead and this cost may rise quickly with the number of correctable errors. For example, a conventional SEC code requires m+1 check-bits for 2m data-bits, while a conventional DEC code needs 2(m+1) check-bits for the same number of data-bits.

In this work, we consider word-oriented binary memories which enable the identification of bits with a reduced level of confidence, called erasures. Such bits are not necessarily erroneous, but they have a higher probability to be affected by errors. Erroneous bits which involve erasures are easier to correct since their positions are known. We assume that the identification of erasures does not require supplementary memory operations. For example, this may be the case of MRAM cells where the information is encoded by the electrical resistance of a magnetic tunnel junction (MTJ). An MTJ may be affected by soft or hard dielectric breakdown which can cause the drifting of its electrical resistance outside the characteristic range of healthy MTJs.

Erasures may also occur in aggressively scaled memories if distinct logic values are encoded by overlapping distributions of some relevant electrical parameter. For example, the electrical charge stored on the floating gates of EEPROM/flash memory cells may be distributed over slightly overlapping ranges. Any read-out value that falls into the overlapping region may be considered as an erasure.

Different approaches to extract and use erasure information are available. When the memory latency and endurance are not an issue, the double-complement algorithm, can be used to detect erasures induced by permanent faults. Soft-decision decoding, usually applied to LDPC codes, offers a natural support to exploit erasure information. Algebraic decoding able to handle erasures has also been proposed. Unfortunately, such complex decoding methods are more suitable for block-oriented low-latency storage devices and rather inappropriate for word-oriented fast memories.

We propose two error correction schemes, illustrated in Fig. 1 and Fig. 2, which rely on erasure information in order to boost the number of errors that can be corrected with SEC-DED codes. Erasure information is only exploited in the presence of detectable multi-bit errors so that single-bit errors are always corrected. Double-bit errors become correctable in the presence of at least one erroneous bit indicated as erasure. Successful decoding in the presence of correct bits indicated as erasures requires that both erroneous bits are identified as erasures. It is shown that virtually all double-bit errors become correctable if the probability that erroneous bits are indicated as erasures is larger than 90%.

These schemes have also been analyzed in combination with shortened SEC codes. Here, an ECC is called shortened if the number of data-bits per code word is below the maximal limit allowed by the available check-bit number. The use of shortened ECCs is imposed by the use of memory word sizes equal to a power of 2. In such cases, important fractions of double-bit and triple-bit errors can be detected with shortened SEC and SEC-DED codes, respectively.

In order to increase the error correction capability of the proposed schemes, ways to construct shortened SEC codes with a high number of detectable double-bit errors and shortened SEC-DED codes with improved triple-bit error detection are presented as well.

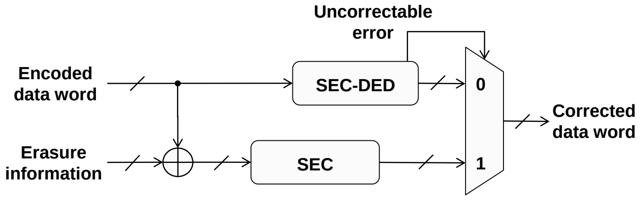


Figure 1 : DEC scheme based on erasure information and conventional SEC and SEC-DED decoders.

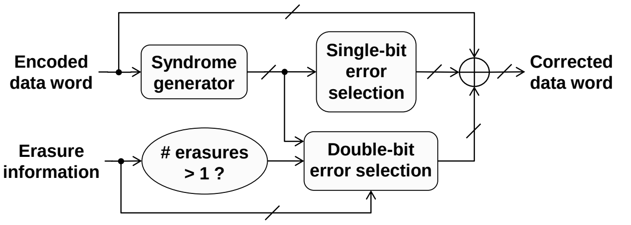


Figure 2 : DEC scheme based on a SEC-DED code, erasure information and an unconventional decoder.

Related Publications:
[1]S. Evain, V. Gherman "Error correction schemes with erasure information for fast memories," IEEE European Test Symposium, pp. 1-6, 2013.

Memory Reliability Improvement based on Maximized Error-Correcting Codes

Research topics : Reliability; MTTF; ECC

V. Gherman, S. Evain, Y. Bonhomme

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

Error-correcting codes (ECC) provide an effective way to achieve the required level of transient fault tolerance in storage and memory subsystems since they can be applied at system or component levels with relatively limited design overhead. Since the implementation of an ECC requires a certain amount of storage overhead, any approach able to boost the fault masking capacity is helpful.

An approach with lower performance overhead is to extend the error correction capability of an ECC without increasing the check-bit number per code-word. This is possible due to the fact that usually the number of data-bits needs to be a power of 2 or a multiple of a power of 2. For example, an (16, 8) ECC allows the correction of all single-bit and double-bit errors, which means 136 errors, while, in principle, up to 255 distinct errors could be distinguished.
Most of these ECC extensions are devoted to the correction of burst errors which affect contiguous codeword bits. This choice is not justified when the burst errors are not necessarily the most probable multi-bit errors. Examples include CMOS memories protected against multi-bit upsets with the help of bit interleaving or non-volatile memories where the corruption of information is not necessarily induced by ionizing particles such as the magnetic RAMs (MRAM).

In this work, we propose the concept of maximized ECCs obtained by a better utilization of the information redundancy available in linear block ECCs. Starting from an N-bit ECC, the goal is to get a maximum number of correctable (N+1)-bit errors. Selection criteria are defined in order to reduce the impact on the latency of the error correcting logic. For example, among all correctable (N+1)-bit errors which generate the same syndrome, the errors which affect a maximum number of check-bits are privileged.
The impact on mean-time-to-failure (MTTF) of a memory was evaluated for the particular cases of ECCs that enable the correction of errors that can affect a maximum of one bit or two bits.

Consider a memory unit protected by an N-bit ECC. Memory failures occur if codewords with more than N corrupted bits are accessed. It is assumed that:
- The (N+1)-bit errors are independent and identically distributed with a given rate λ,
- The rate of errors affecting more than N+1 bits is much smaller than λ.

The probability P that a memory unit operates without failure during a time interval τ can be expressed as:

$$P = e^{-\lambda \cdot \tau}$$

where $1/\lambda$ gives the memory MTTF under the specified assumptions.
If a certain fraction x ($0 \le x \le 1$) of the (N+1)-bit errors can be masked, the (N+1)-bit error rate responsible for memory failures will become $(1-x)\cdot\lambda$ and the memory MTTF improvement is given by the following expression:

$$\frac{\Delta MTTF}{MTTF} = \frac{x}{1-x}$$

We generate H-matrices of the maximized double-bit error correcting codes (DEC) from scratch with the help of a SAT-solver. If only data-bits need to be provided by a maximized DEC decoder, the hardware overhead can be reduced by selecting those triplets that involve a maximum number of check-bit positions. Besides the constraints specific to a DEC code, two additional goals were imposed:
- Find an H-matrix that can provide a maximum number of triple-bit errors that can be corrected,
- Maximize the number of correctable triple-bit errors that involve only check-bit positions.

Once the H-matrix is found, the set of correctable triple-bit errors is constructed by selecting first those triple-bit errors which involve only check-bit positions, then those which involve two check-bit positions, and finally those which involve one or zero check-bit positions.
Table 1 reports the number of triple-bit errors that can be masked with the fastest conventional DEC decoders and the achieved memory MTTF improvement enabled by the maximized DEC decoders. A memory MTTF improvement between 20% and 30% was achieved.

Table 1: MTTF improvement with the maximized DEC decoders

DEC code	Number of triple-bit errors that can be masked with the fastest DEC decoder	Achieved number of correctable triple-bit errors with the maximized DEC decoder	Achieved MTTF improvement
(16,8)	29	118	20%
(21,12)	36	280	23%
(26,16)	92	672	30%
(35,24)	73	1417	26%
(44,32)	102	3100	30%

Related Publications:
[1] V. Gherman, S. Evain, Y. Bonhomme, "Memory Reliability Improvement Based on Maximized Error-Correcting Codes", Journal of Electronic Testing 29(4), 601-608, 2013.

Scan design with shadow flip-flops for low performance overhead and concurrent delay fault detection

Research topics: shadow scan; monitoring; concurrent fault detection

S. Sarrazin, S. Evain, L. Alves de Barros Naviner, Y. Bonhomme, V. Gherman

ABSTRACT: A shadow-scan solution is proposed in order to provide on-line monitoring support and reduce the latency overhead. This is achieved with the help of a shadow scan flip-flop (FF) which is associated to a system FF. During test, system observability is naturally ensured by the shadow FF and system controllability is enabled with the help of asynchronous set and reset operations applied to the system FF. An set operation is executed in scan mode and a selective reset operation is executed during the transition phase between scan and capture modes. It is shown that a lower latency overhead can be achieved as compared to standard scan design.

Energy efficiency, usually measured as performance per watt, is an important figure of merit for mobile devices. Dynamic voltage and frequency scaling (DVFS) is an effective solution to save energy based on dynamic adjustments of the supply voltage and clock frequency. Such adjustments are limited by the necessity to introduce temporal safety margins to absorb dynamic latency variations that may be induced by wear-out or voltage and temperature variations.

On-line monitoring based on concurrent delay fault detection is a natural approach to reduce safety margins for aggressive DVFS. Here, we consider a concurrent fault detection scheme in which a shadow flip-flop (FF) is associated to each system FF placed at the end of timing-critical paths. In order to enable error prediction, the shadow FFs must have a higher probability to be affected by DVFS-induced delay faults than the system FFs. One way to achieve this is to select shadow FFs with a larger setup time than system FFs. Care should be taken since the additional setup time may affect circuit latency.

On the other hand, product quality requires the execution of high quality manufacturing tests and diagnosis. Standard scan is a widely employed technique to effectively address these issues. Unfortunately, standard scan design is also a source of latency overhead since the system FFs need to be replaced with slower scan FFs.

Here, shadow scan solutions are proposed to reduce the latency overhead of scan and concurrent delay fault detection. This solution is applied to system FFs located at the end of timing-critical paths in order to avoid their scan chain insertion. Mixed scan architecture may be obtained if standard scan design is applied to the remaining FFs as illustrated in Fig. 1.

During test, shadow scan design does not affect system observability since test responses can still be captured and scanned out. Full controllability can be ensured if the system FFs with shadow scan are set in scan mode and selectively reset before the system switches to capture mode. The set operation can be controlled by the scan enable signal while the selective reset operation is controlled by the values scanned into the shadow FFs.

Fig. 2 details a shadow scan design which enables concurrent delay fault detection during system mission and system observability and controllability during test. This design also provides concurrent detection of soft errors induced by transient faults which affect the system FF, the shadow FF or the upstream combinational logic. To that end, the xor-gate output should be interpreted as an error and not as a warning signal.

It is shown that shadow scan design with asynchronous set and reset may have a lower latency overhead than standard scan design. For example, in the TSMC N40LP standard-cell library, the FFs with asynchronous set and reset have a setup time between 3 and 4 times smaller than in the case of standard scan FFs with or without asynchronous reset.

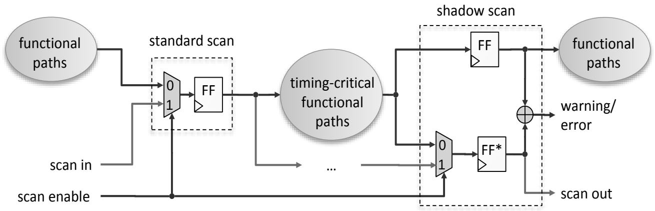


Figure 1 : Mixed scan architecture with standard and shadow scan design.

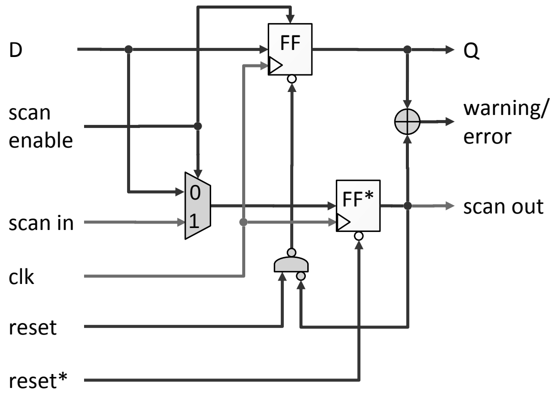


Figure 2 : Fully controllable system FF with shadow scan FF (FF) that enables concurrent delay fault detection.*

Related Publications:
[1] S. Sarrazin, S. Evain, L. Alves de Barros Naviner, Y. Bonhomme, V. Gherman "Scan design with shadow flip-flops for low performance overhead and concurrent delay fault detection," IEEE Design and Test in Europe (DATE), pp. 1077–1082, 2013.

Time- and angle-triggered real-time kernel and its use for Powertrain applications

Research topics: Time-Triggered paradigm, Real-Time Operating System

D. Chabrol (Krono-Safe), D. Roux (Krono-Safe), V. David, M. Jan, M. Hmid, P. Oudin (Delphi) et al.

ABSTRACT: The time-triggered (TT) approach provides a predictable and reproducible execution of real-time systems but cannot cope with tight temporal constraints (around 100µsec) and does not allow to directly specifying the temporal behavior of the system based on angles. We have extended the PharOS technology to combine several time domains (time and angle triggered) allows designing and executing powertrain controllers in a deterministic way on multi-core architectures. We have developed a prototype of a subset of a powertrain controller from Delphi based on PharOS.

Embedded safety-critical real-time systems are often implemented as hard real-time periodic tasks within a Time-Triggered (TT) real-time operating system (RTOS). The schedulability analysis of such systems must be performed assuming the worst-case demand. However, instrumentation and control multitasking systems can have tight temporal requirements. Besides, the physical law of the system to control and command may dynamically change within a specified range. The use of the TT paradigm to design such systems dramatically oversizes the required processing capability, if not making systems unschedulable.

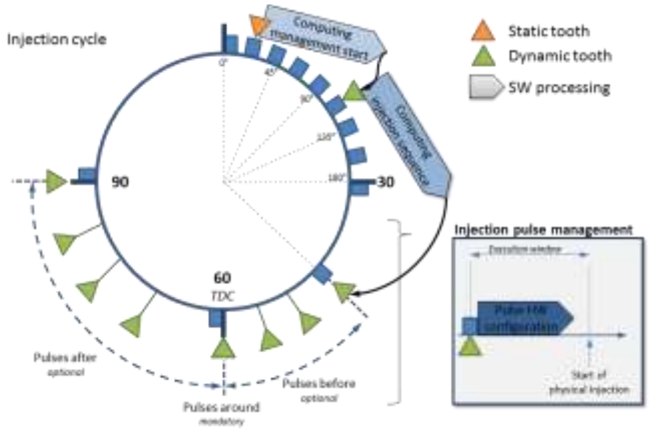


Figure 1. Injection cycle in a powertrain use case made of a sequence of pulses on tooth dynamically computed. Computations start on a static tooth and are adjusted dynamically.

In the automotive field, the management of an engine is ensured by a powertrain controller. It is an electronic device that assists the control of vehicle's engine by measuring multiple events and enabling real time adjustments of fuel, air, spark and gear shift to help the powertrain system to efficiently, reliably and economically operate. Its main function is to compute when the injection sequences of fuel in the motor should occur (as shown by Figure 1). The frequency of injection depends on the speed of the engine. In order to reduce of clatter of the motor as well the gas emissions and the fuel consumption, the injection sequence should occur at precise instants. However, the execution time on an injection sequence is independent of the engine speed. As the speed of engine may vary, the starting point of an injection sequence should then be dynamically adjusted in order to be the most accurate. This can lead to temporal of a few hundred microseconds.

We have proposed a new paradigm, called eXternal-Triggered (xT), which generalizes the TT approach to external events and therefore unifies this model to the Event-Triggered paradigm. The triggering of tasks can be linked to the occurrence of events from the controlled system, in addition to the classical physical time scale of the TT paradigm. Input events, such as temporally mastered interrupts, are used by the xT paradigm to produces ticks available for specifying the triggering of tasks. However, each event may not be linked to a tick: several events may be required to produce a tick. This feature is implemented by defining a filtering function for each source of input events. It must be provided by the designer of a domain as it knows the meaning of events from the controlled system. It can also be used to freeze the current time of a domain by filtering any event that will occur. This can be used to implement a synchronization strategy between the value of the current time of a domain and the physical process being control and command or simply in case of hardware failure of the sensors that generate events.

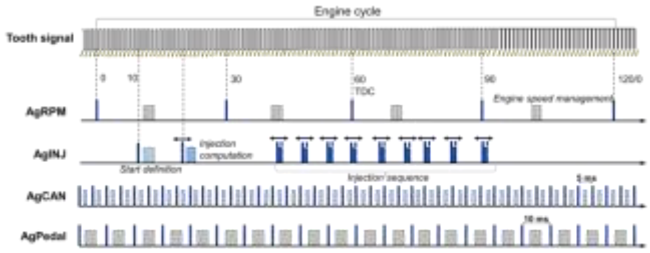


Figure 2. Engine controller temporal diagram.

We have extended our PharOS Real-Time Operating System (RTOS) technology to include the support of the xT paradigm and we applied it to develop of a subset of a powertrain controller from Delphi over a PowerPC MPC551x evaluation board. Figure 2 shows the temporal behavior of the designed application.

Related Publications:

[1] Damien Chabrol, Didier Roux, Vincent David, Mathieu Jan, Moha Ait Hmid, Patrice Oudin, Gilles Zeppa, "Time- and angle-triggered real-time kernel". Proc. of the Design, Automation and Test in Europe (DATE2013), pages 1060-1062, March, Grenoble, France.

Optimizing the use of multicores real-time systems: the elastic task model for mixed-criticality and cache-aware scheduling

Research topics: real-time multicore scheduling, cache hierarchy, mixed-criticality

Mathieu Jan, Lilia Zaourar, Maurice Pitel (Schneider Electric Industries)

ABSTRACT: industrial fields must build at the most competitive price real-time systems made of an increasing number of functionalities. This can be achieved by hosting high-criticality tasks as well as consumer real-time low-criticality tasks on a same chip, leading to the design Mixed-Criticality (MC) systems, but also by enhancing the scheduling being using. We introduce a variant of the elastic task model for maximizing the execution rate of the low-criticality tasks and propose a formulation for computing a static scheduling that minimize L1 data cache misses between hard real-time tasks on a multicore architecture using communication affinities.

Traditionally, industrial systems use a dedicated chip for executing a set of real-time tasks with a same level of criticality. When such tasks are safety-critical, high margins are taken on their Worst-Case Execution Time (WCET). This leads to the specification of high allocated budgets of time for such high-criticality tasks. Besides, the probability that the WCET of a set of high-criticality tasks occur simultaneously is very low. However, the schedulability demonstration of safety-critical systems must be performed in the worst-case situation, due to certification constraints. This therefore leads to a huge over-sizing of the CPU resources that are needed compared to what is really used, in average, while the system is running. This practice becomes incompatible with the current trend of tighter economical constraints of various industrial domains, such as the automotive or energy distribution fields. Therefore, there is a need to use these generally unused processing capabilities for executing low-criticality tasks. This need is further exacerbated when the underlying hardware is a multicore architecture.

The problem in such mixed-criticality systems (MC) is to increase the schedulability of the low-criticality tasks, while still guaranteeing in the worst-case scenario the schedulability of the high-criticality tasks. From the existing solutions, the elastic task model avoids degrading the service level when a low-criticality task misses a deadline. The first contribution of our work [1] is the definition of a variant of an elastic task model for maximizing the utilization of the processing capabilities of an architecture, in which high- and low-criticality tasks are schedulable taken separately, while the sum is not. Off-line, we use a linear programming approach to compute the different stretching factors that must be applied on the periodicity of the low-criticality tasks, so that the schedulability of the high-criticality tasks are guaranteed. On-line, we bet on the availability of slack time generated by high and low-criticality tasks. No stretching factors are therefore applied in order to execute the low-criticality tasks at their fastest rates. However, when a deadline is going to be missed by a low-criticality task, its deadline is stretched up to point that prevents the deadline miss. Figure 1 shows the distribution of such stretching factors between two different configurations for a same task set.

In configuration A, random values for the importance level, while in configuration B the stretching factor or some tasks is set to 1.25. Such control over the values of the stretching factor for each task opens the opportunity for application designers to more easily dimension the required processing power when designing a MC system.

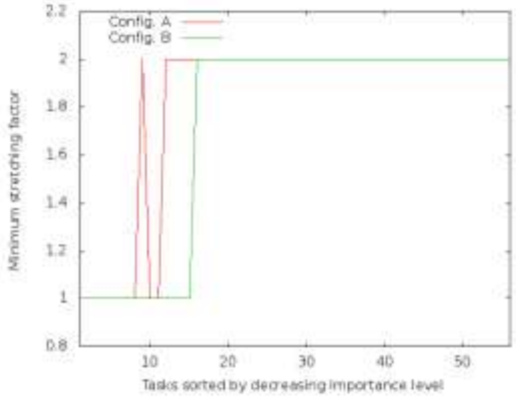


Figure 1: evolution of the distribution of stretching factors when setting them to a fixed value for some low-criticality tasks.

We also show how our proposed approach to deal with the low-criticality tasks can be used within the Time-Triggered paradigm, used to build safety-critical real-time systems. When using the elastic task model, the low-criticality tasks have defined triggering points assuming the non-stretched temporal behavior. This leads to an inconsistency between the expected temporal behaviors of the tasks, if the stretching factor of a single task is modified. We propose a decision algorithm for appropriately setting stretching factors of tasks in these cases.

The growing need for continuous processing capabilities has also led to the development of multicore systems with a complex cache hierarchy. Such multicore systems are generally designed for improving the performance in average case, while hard real-time systems must consider worst-case scenarios. An open challenge is therefore to efficiently schedule hard real-time tasks on a multicore architecture. We have propose a mathematical formulation for computing a static scheduling that minimize L1 data cache misses between hard real-time tasks on a multicore architecture using communication affinities [2].

Related Publications:
[1] M. Jan, L. Zaourar, M. Pitel, "Maximizing the execution rate of low criticality tasks in mixed criticality system", Proc. of the First Workshop on Mixed-Criticality Systems (WMC), pages 43-48, December 2013, Vancouver, Canada.
[2] L. Zaourar, M. Jan, M. Pitel, "Cache-aware static scheduling for hard real-time multicore systems based on communication affinities", Proc. of the WiP session of the 34th IEEE Real-Time Systems Symposium (RTSS'13), pages 3-4, December 2013, Vancouver, Canada.

Scheduling Algorithms to Reduce the Static Energy Consumption of Real-Time Systems

Research topics: real-time multiprocessor/multicore scheduling, energy consumption

Vincent Legout, Mathieu Jan, Laurent Pautet (Telecom ParisTech)

ABSTRACT: energy consumption of real-time embedded systems is a growing concern. It includes both static and dynamic consumption and is now dominated by static consumption as the semiconductor technology moves to deep sub-micron scale. In these work, we propose a new approach to efficiently use the low-power states of multiprocessor embedded real-time systems in order to reduce their static consumption. We have targeted both hard real-time and mixed-criticality systems, where we exploit the ability of tasks with low-criticality levels to cope with deadline misses.

Real-time embedded systems tend to have a limited power supply. Therefore minimizing their energy consumption is an important concern, for instance, in the automotive and the energy distribution fields. The energy consumption can be divided into two categories: dynamic and static consumption. Dynamic consumption depends on the activity of the processors. On the other hand, static consumption is mainly due to leakage current and is present even when no operations are performed. Dynamic consumption used to dominate static consumption for micrometer-scale semiconductor technology. The leakage current already accounted for 50% of the total power dissipation in 90 nm technologies and this trend is only increasing as the VLSI technology is scaling down to deep sub-micron domain.

Industrials designers of safety-critical systems subject to certification constraints are slowly considering multiprocessors for their next generation of systems. Due to economical constraints and thanks to this increased processing capability, designers also aim to execute applications with different levels of criticality on a same multiprocessor chip. Reducing the energy consumption of mixed criticality systems while enforcing their schedulability is yet an open research topic.

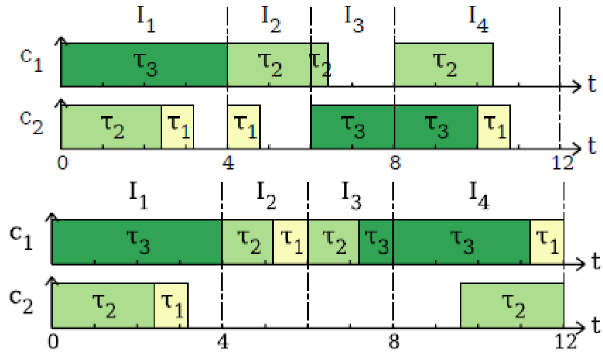


Figure 1: scheduling without (top) and with our LPDPM algorithm (bottom) on a 2 processors system (C1 & C2).

The first contribution of our work is, to the best of our knowledge, the first global optimal multiprocessor scheduling algorithm reducing static consumption [3]. Our algorithm is called LPDPM for Linear Programming DPM and generates a schedule guaranteeing the schedulability of the task set and minimizing the static energy consumption. To this end, we produce fewer but longer idle periods as shown by Figure 1. It further extends the length of the idle periods, especially when tasks do not consume all their Worst Case Execution Time (WCET). When tasks use their WCET, the energy consumption of LPDPM while processors are not executing tasks is up to 8 times smaller than with recently proposed optimal state-of-the-art multiprocessor schedulers (RUN & U-EDF), as shown by Figure 2.

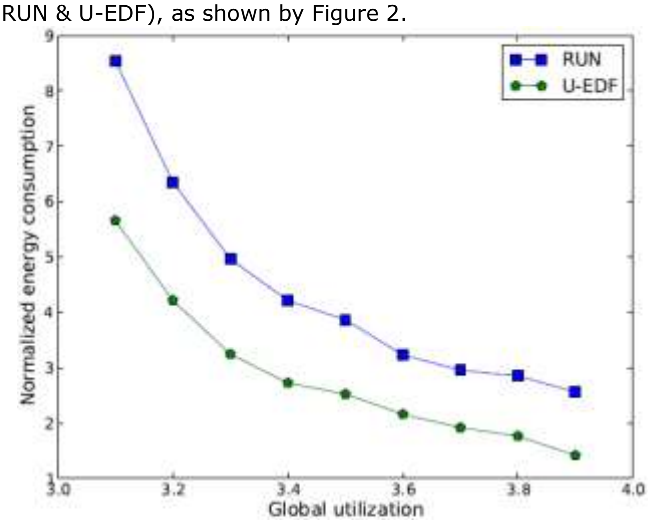


Figure 2: overall idle consumption of RUN & U-EDF compared to LPDPM (energy consumption of LPDPM=1).

Our second contribution [2] exploits the ability of tasks with low-criticality levels to cope with deadline misses by being more aggressive while scheduling the low-criticality tasks. The rationale is that not guaranteeing offline the deadlines of all the low-criticality tasks is not a major issue, as the tasks (low and high-criticality) often do not use WCET. The generated slack time can therefore be used on-line by the other low-criticality tasks to meet their deadlines. Simulations show that using the best compromise, the energy consumption can be reduced up to 17% while the percentage of deadline misses is kept under 4%.

Related Publications:
[1] V. Legout, M. Jan, L. Pautet, "An off-line multiprocessor real-time scheduling algorithm to reduce static energy consumption", Proc. of the First Workshop on Highly-Reliable Power-Efficient Embedded Designs (HARSH), pages 7-12, February 2013, Shenzhen, China.
[2] V. Legout, M. Jan, L. Pautet, "Mixed-criticality multiprocessor real-time systems: Energy consumption vs deadline misses", Proc. of first Workshop on Real-Time Mixed Criticality Systems (ReTiMiCS), pages 1-6, August 2013, Taipei, Taïwan.
[3] V. Legout, M. Jan, L. Pautet, "A scheduling algorithm to reduce the static energy consumption of multiprocessor real-time systems", Proc. of the 21st Intl. conf. on Real-Time Networks and Systems (RTNS), ACM, pages 99-108, October 2013, Sophia Antipolis, France.

Divide and Conquer: Separation of Concerns to Parallelize Embedded System Developments

Research topics : Embedded System, Development Method, Team Work

N. Hili, C. Fabre, S. Dupuy-Chessa (LIG, SIGMA), D. Rieu (LIG, SIGMA)

ABSTRACT: Embedded System (ES) development distinguishes itself from both computer science and computer engineering and requires specific approaches to organize developments in a productive fashion. Besides this initial scientific complexity several industrial aspects of ES engineering increase further development complexity: Strong market pressures on product costs (bill of materials), time to market and development costs. We propose ⟨HOE⟩2, a method dedicated to embedded system development that addresses both engineering aspects and industrial concerns related to efficient work organization.

One of the most challenging aspects of Embedded System (ES) development is the heterogeneity of fields and skills required during design, as illustrated in Fig. 1. Each field has its own view of the system under design and they all need to converge in the final system. In particular, as ES design integrates both software and hardware into a homogeneous solution it requires deep collaboration of the stakeholders. As the discipline is young, methods and tools for embedded system design lack of maturity, hindering the broad-scale development.

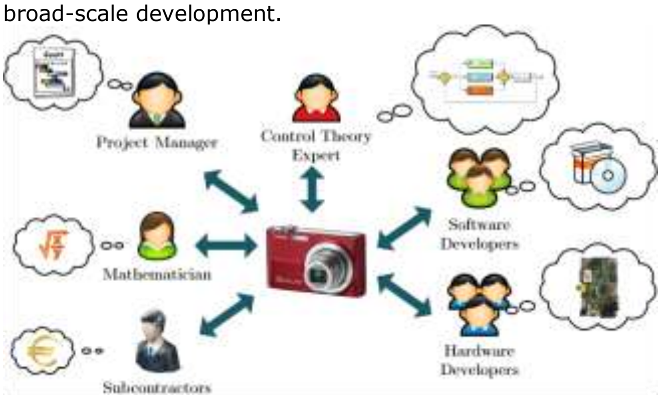


Figure 1: Stakeholder's View for an Embedded System

Several advances in platform-based and model-based development contribute to reduce this complexity. One of them is the application of the Model-Based Development (MDD) for embedded system development and transformations of platform-independent models to platform-specific to perform code generation.

In [1], we presented ⟨HOE⟩2 (for "Highly Heterogeneous Object-Oriented Efficient Engineering"), a model-based method dedicated to embedded system development. ⟨HOE⟩2 proposes a platform-based strategy to ease ES development. We extended this strategy to the platform composition aided in a fractal process (see Fig. 2). Platform composition allows considering complex platform development as a result of the composition of simpler ones, to build the desired solution by refinement, increasing significantly the opportunities of parallel developments and team organization.

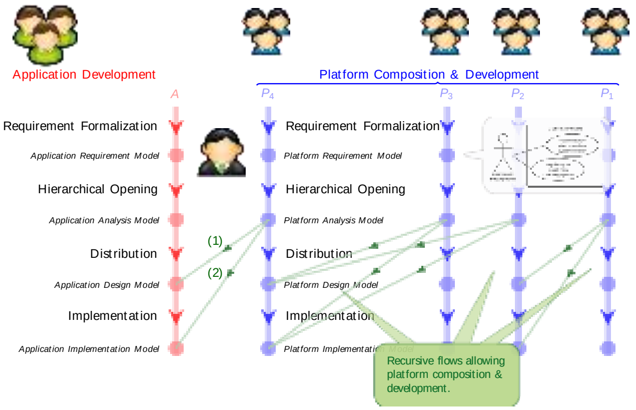

Application Development
Platform Composition & Development
A
P_4
P_3
P_2
P_1
Requirement Formalization
Application Requirement Model
Hierarchical Opening
Application Analysis Model
Distribution
Application Design Model
Implementation
Application Implementation Model
(1)
(2)
Requirement Formalization
Platform Requirement Model
Hierarchical Opening
Platform Analysis Model
Distribution
Platform Design Model
Implementation
Platform Implementati
Recursive flows allowing platform composition & development.

Figure 2: The ⟨HOE⟩2 Fractal Process

The tooling of ⟨HOE⟩2 called for graphical modeling tools. Embedded system designers increasingly use graphical modeling tools. Both open source and commercial tools proposes different approaches to design graphical modeling editors. In [2], we experimented and compared a number of existing graphical modeling tools and we demonstrated how Graphiti (http://www.eclipse.org/graphiti) can be used to efficiently and rapidly build graphical modeling editors through three video-based tutorials.

We decided to pick Graphiti to develop a dedicated tool for ⟨HOE⟩2. The resulting tool is presented in Fig. 3.

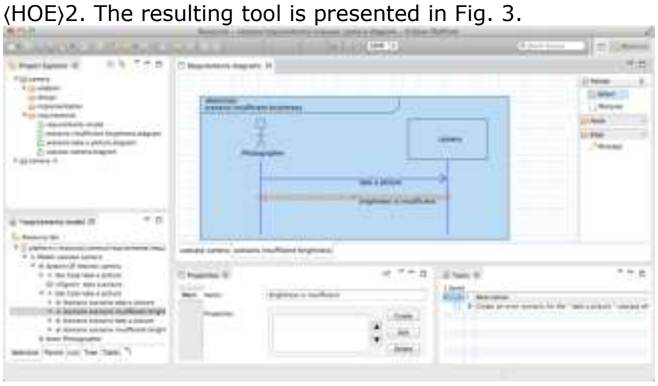


Figure 3: CanHOE2, a Dedicated Tool for ⟨HOE⟩2

Related Publications :
[1] N. Hili, S. Dupuy-Chessa, D. Rieu, C. Fabre, "Divide and Conquer: Separation of Concerns to Parallelize Developments", Summer School on Cyber-Physical Systems, Jun. 2013.
[2] N. Hili, "How to Efficiently and Rapidly Build Graphical Modelling Editors using Graphiti" (Manual), Eclipse DemoCamps, Nov. 2013.

An efficient parallelization strategy for dynamic programming on GPU

Research topics : Dynamic programming, GPU computing, combinatorial optimization.

K.-E. Berger, F. Galea

Optimization methods generally do not fall into the most suitable algorithms for parallelization on a GPU. However, a relatively good efficiency can be obtained if the method is properly adapted to the GPU programming model. In our work, we propose a parallelization strategy for thread grouping for dynamic programming in Nvidia CUDA. This strategy provides good acceleration compared to a standard GPU parallel strategy on a dynamic programming-based implementation of the knapsack problem. We show this strategy is helpful in the case of the multi-dimensional knapsack problem, where computing multi-dimensional indices is a costly operation.

Since the introduction of programmable Graphics Processing Units (GPUs), the scientific community quickly understood the interest in using their computing power for scientific computations. Even though those processors were firstly designed for graphics rendering, they provide a much higher computing power than the general purpose Central Processing Units (CPUs) in these computers.

So called GPGPU (General Purpose GPU) are difficult to exploit in order to get an efficient use of their computation power. This is due to the complexity of their specific architecture (e.g. Figure 1). Despite of this difficulty, GPUs, due to their massive internal parallelism, can provide a good acceleration for computations if they are programmed properly, and if the computations to be performed expose a significant enough level of parallelism.

We take the multidimensional knapsack problem (MKP) as a use case. MKP is a multidimensional extension of the well-known knapsack problem (KP). Even though large instances of KP and MKP are usually solved to near optimality using collaborative approaches of metaheuristic and Integer Linear Programming methods, optimal solving of small instances can be efficiently performed using dynamic programming (DP).

This work focuses on an efficient parallelization of the general DP process on GPU, which can be used for parallelizing any DP-based algorithm. This is especially useful when the problems to be solved are small subproblems of a larger general problem. In this case, it may be relevant to reduce each step's computation time as much as possible; GPU parallelization is a possible approach to achieve this goal.

The first developed strategy is based on a particular interpretation of the Nvidia CUDA programming documentation, which emphasizes the need to maximize the parallel execution. The commonly observed behavior in GPGPU computing for achieving this goal is to spawn as many threads as there are independent parallel operations. Thus, this method consists in executing the sequential computing on CPU and calling the GPU as many times as required with the finest grain for parallel computing. The necessary synchronization of all GPU threads is implicit as the CPU waits for all GPU threads to finish their execution. This is the common strategy used in many related works in parallel dynamic programming on GPU.

A second developed strategy results from the observation that each thread performs a very little amount of work. This led us to investigate on limiting the number of spawned threads while maximizing the GPU utilization and minimizing the intervention of the CPU in the algorithm. It consists in implementing the whole KP algorithm in a GPU kernel.

While thread synchronization is supported by CUDA for threads within the same block, there is no supported way for synchronization for threads from different blocks due to the GPGPU restrictions.

However, we implemented a barrier mechanism between parallel running blocks, maximizing the number of blocks running in parallel with regards to the hardware specification of the GPU.

The final developed strategy is an intermediate solution which follows the scheme of the first strategy, while reducing the number of blocks at the same level as the second strategy. There is no longer the need for thread synchronization, as we get similar implicit synchronization as the first strategy. This parallel strategy produces very similar results as the second strategy, while its design is much simpler and does not require undocumented tricks to achieve its goal. This is the strategy we proposed in [1].

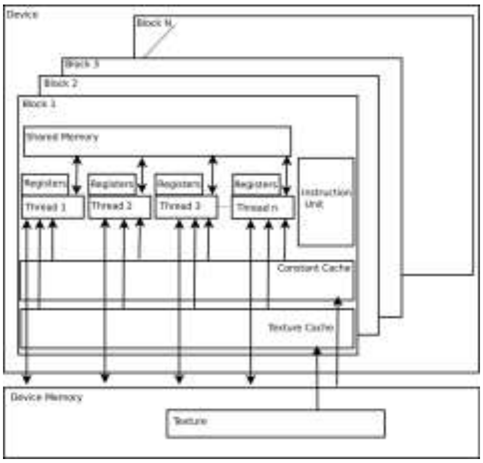


Figure 1 : Memory structure in a GPU device

Related Publications:
[1] K.-E. Berger, F. Galea, "An efficient parallelization strategy for dynamic programming on GPU", 2013, IEEE 27th International Symposium on Parallel & Distributed Processing Workshops and PhD Forum, PCO.

Throughput constrained parallelism reduction in cyclo-static dataflow applications

Research topics : parallelism reduction, dataflow programming, CSDF, actor fusion

S. Carpov, L. Cudennec, R. Sirdey

ABSTRACT: This paper deals with semantics-preserving parallelism reduction methods for cyclo-static dataflow applications. Parallelism reduction is the process of equivalent actors fusioning. The principal objectives of parallelism reduction are to decrease the memory footprint of an application and to increase its execution performance. We focus on parallelism reduction methodologies constrained by application throughput. A generic parallelism reduction methodology is introduced. Experimental results are provided for asserting the performance of the proposed method.

Nowadays, much effort is dedicated to the study of many-core computing systems, beginning with hardware architecture design issues and ending with software programmability aspects. The main difficulty of efficient utilization of parallel systems resides in their programming, both in terms of conception time and as well as execution performance. The dataflow model of computation has been purposely introduced to facilitate parallel systems programming.

A dataflow application is a network of actors communicating through unbounded, unidirectional FIFO channels and exclusively through these channels. One instantiation is the cyclo-static dataflow (CSDF) graph. CSDF model is particularly well suited for programming embedded systems because several important application properties (absence of deadlock, bounded memory execution etc.) can be proven. The main goal of this study is to obtain a new application A' which is semantically equivalent to the initial application but with fewer “parallelism” in it. We call this action parallelism reduction. Our work is particularly aimed at parallelism reduction in ΣC applications.

The advantages of parallelism reduction are: memory footprint of application binaries decreases (less redundancy in code/data loading), program compilation is faster, scheduling overhead is lower and by consequence system times are smaller etc. These advantages are even more important in embedded systems where on-chip memory size is small and scheduling algorithms are sensible to the number of actors. The sizing of applications in embedded systems has to meet the following requirements: (i) being parallel enough in order to offer the desired application throughput and (ii) being small enough in order to fit the memory footprint of the target chip.

The parallelism reduction problem is not well known to the literature. One can mention the paper [1] to which our work resembles the most. The authors describe a pattern substitution based method for parallelism reduction in ΣC [2] applications. Their goal is to bound the number of actors per processing core to a predefined limit. While reducing the memory footprint, this approach does not ensure that the execution throughput is preserved.

In this work we introduce a generic parallelism reduction method. The proposed method does not depend on a predefined set of patterns and is not limited to horizontal or vertical actor fusion as in StreamIt language. It reduces the inherent application parallelism in function of actor execution times and application throughput constraints.

In ΣC applications instances of the same actor are called equivalent actors. Equivalent actors perform the same computation but on different data streams. Two or more equivalent actors can be merged together. The corresponding input (output) data streams are time-(de)multiplexed using join (split) actors. One way to interpret actor fusion is that the execution of actors is serialized.

Initially application actors are partitioned into sets of equivalent actors, i.e. instances of the same actor are grouped together. The sets containing only one actor are discarded directly. Each set is partitioned into subsets such that their fusion respects the throughput constraint and afterwards the actors of each subset are fusioned together. The partitioning of equivalent actors and the order of actor fusion are potential optimization goals [3].

In Figure 1 is shown the cumulated execution time for 50 runs of a LoG edge detection algorithm without and with two types of parallelism reduction. As we can see the parallelism reduction decreases by 30% the execution time.

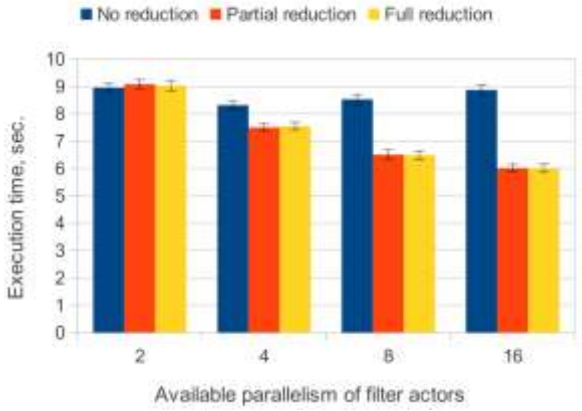


Figure 1: Execution time decrease for LoG edge detection algorithm.

Related Publications :

[1] L. Cudennec, R. Sirdey, "Parallelism reduction based on pattern substitution in dataflow oriented programming languages", 12th International Conference on Computational Science, 2012.

[2] T. Goubier, R. Sirdey, S. Louise, V. David, "ΣC: A Programming Model and Language for Embedded Manycores", Algorithms and Architectures for Parallel Processing, 2011.

[3] S. Carpov, J. Carlier, D. Nace, R. Sirdey, "Task ordering and memory management problem for degree of parallelism estimation", Lecture Notes in Computer Science, Vol. 6842, 2011.

Automatic deployment on embedded parallel systems

Research topics: parallelism, automation, adequacy, exploration, benchmarking

M. Krumpe Goldsztejn, Y. Lhuillier, J. Falcou (LRI), L. Lacassagne (LRI)

In order to be able to let developers write efficient and portable programs, we propose a programming approach based on generative programming and algorithmic skeletons.
In our approach, generative programming (or metaprogramming) is used to separate architectural specific constructs from algorithmic description.
A single algorithm description allows our tool to generate the deployment for multiple architectures, and even multiple topologies for a same target.

Although Parallel architectures have being studied for several decades the trend is still the multiplication and diversification of these architectures which also applies to their programming models. This diversity is particularly remarkable in embedded systems that are designed for specific applications. Such architectures have various power and size constraints that make parallel embedded architectures highly specialized, heterogeneous and require specific programming. This not only means complexity to program but also complicates comparing of these systems. Moreover the programmers need to learn a new programming model for every architecture.

In order to address both the issues of portability and benchmarking we have designed a tool implementing an automatic methodology for the deployment of code [1]. We propose a programming approach based on generative programming and algorithmic skeletons. The structure of our solution is presented in Fig.1 and consists of 3 levels.

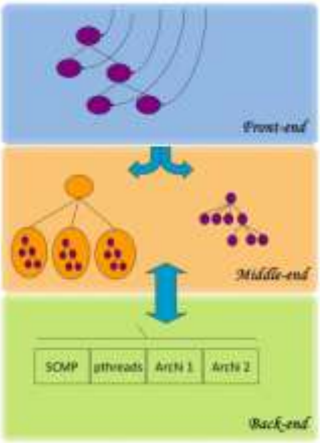


Figure 1: Overall structure of the deployment methodology

The front-end is based on an algorithmic skeletons approach: the user writes a skeleton description of the application which is translated into an internal representation. At the middle-end all possible semantics is extracted using architecture agnostic transformations of the internal representation. Finally the back-end produces target architecture specific files required for compilation.

The CEDAR Architecture (Fig. 2) is a Configurable Embedded Distributed Architecture with an adaptive routing strategy based on ACO (Ant Colony Optimization).

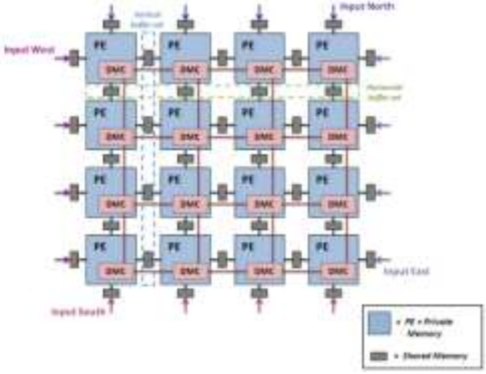


Figure 2: CEDAR architecture

It offers a high degree of flexibility and can handle any interconnection topology (Fig. 3). Routing paths for remote data transfers are defined at runtime and allow a homogeneous distribution of traffic, avoiding deadlocks and contentions.

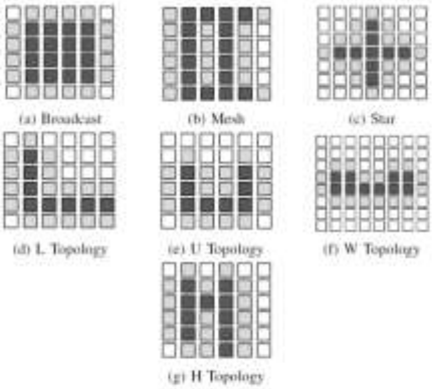

(a) Broadcast
(b) Mesh
(c) Star
(d) L Topology
(e) U Topology
(f) W Topology
(g) H Topology

Figure 3: Topologies of placement for CEDAR

The first experiments of automatic deployment with a CEDAR back-end [3] have proven that most specific codes can be generated using our approach. It is also possible to generate multiple topologies without re-writing any code.

Related publications
[1] M. Krumpe Goldsztejn, Y. Lhuillier, J. Falcou and L. Lacassagne, “Automatic deployment on embedded parallel systems” GDR GPL – CIEL 2012
[2] M. Krumpe Goldsztejn, Y. Lhuillier, J. Falcou (LRI), L. Lacassagne (LRI), "Déploiement automatique de code sur une architecture parallèle embarquée", Rencontres francophones de Parallélisme à la Conférence d’informatique en Parallélisme, Architecture et Système (RenPar à ComPAS), 2013.
[3] C. Azar, S. Chevobbe, Y. Lhuillier, J-P Diguet, "Dynamic routing strategy for embedded distributed architectures," 18th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2011, pp.653,656

Runtime Code Generation for Constrained Embedded Devices. Application to micro-controllers with less than 1k byte of RAM

Research topics: code generation, micro-controller, sensor network

D. Couroussé, A. Aracil, H.P. Charles, V. Lomüller

ABSTRACT: We achieved runtime code generation in less than 1k byte of RAM. Our technique is applicable on highly constrained devices such as the micro-controllers used in Wireless Sensor Networks. The aim is to provide an application the ability to specialize its application code according to runtime constraints, in order to reduce the energy of computations. We illustrate that it is possible to achieve above 20x speedups compared to conventional approaches.

Computing units embedded in the nodes of Wireless Sensor Networks (WSN) are usually small micro-controllers with very limited computing capabilities and memory resources. They are not designed for intensive data processing, and lack hardware support for computing: micro-controllers usually lack an integer divider; some even lack an integer multiplier. The price and energy costs of a floating-point unit (FPU) are usually not affordable for such platforms. However, considering the energy costs of onboard computation versus the energy costs of sending the data over the network for remote processing, it is still desirable to achieve data processing on board to reduce the overall energy budget of the network, and to increase the autonomy of the nodes. Some WSN applications even require onboard floating-point capabilities even if the lack of a dedicated FPU will incur a very high computational cost.

We observed that, in WSN applications at least, a lot of onboard processing requires the use of coefficients that are likely to change at runtime if the application is reconfigured, but that can be considered as constant in a reduced time frame of the application life. By generating specialized versions of the application code on the basis of the values of these "runtime constants", it is possible to generate processing kernels with better performance. The results presented here have been applied to floating-point multiplication, but are also applicable to a lot of processing patterns used in WSN applications: for example means, scaling values from ADC or to DAC, and in general data filtering.

We develop a tool, named deGoal [1], that gives to an application the capability to tune itself, at runtime, according to the characteristics of the execution context, in particular according to the data to process. The general idea is to embed ad hoc runtime code generators, called "compilettes", in the application code. Compilettes offer a very low memory footprint and a high code generation speed compared to traditional techniques for dynamic compilation. As a matter of fact, micro-controllers are out of reach of traditional dynamic compilation approaches, but we show here that our approach for runtime code generation is achievable on these platforms. Our results were measured on Texas Instruments' Launchpad board, which embeds the MSP430, a 16-bit micro-controller with only 512 bytes of RAM. The results are further detailed in [2].

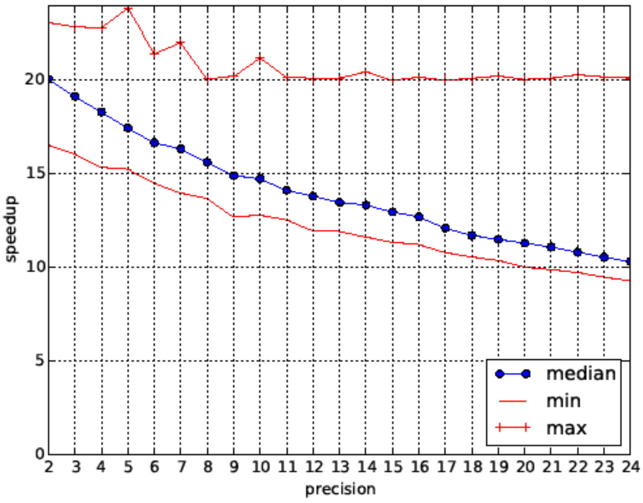


Figure 1: speedup of a 32-bit floating-point multiplication routine generated at runtime, compared with the version of the platform's compiler msp430-gcc. Results are obtained from 4000 samples, with a variable precision in the range [2; 24]. The reference has a fixed precision of 24.

Our specialized routines are generated at runtime and optimized according to the value of one of the two operands. We compare their performance with the hand-optimized versions that come with the platform's compiler msp430-gcc compiled with the full optimisation flags.

Figure 1 presents the speedup results achieved for floating-point multiplication. While the standard version of gcc has a fixed precision of 24 for 32-bits floating-point (i.e. the 24 bits of the mantissa operands are considered for processing), it is possible to further accelerate the computation by varying the precision between 2 and 24, according to the application requirements. Figure 1 shows that with the same precision as used in gcc (24), we achieve speedups between 9x and 20x depending on the data values. Lowering the precision allows further performance improvements. The overhead of code generation is recovered only after 4 calls of the generated code, in the worst case. On a real case application (IIR filter with 32-bits floating-point) we achieved average speedups of 52% [2].

Related Publications:
[1] Couroussé, D.; Lomüller, V. & Charles, H.-P. Introduction to Dynamic Code Generation - an Experiment with Matrix Multiplication for the STHORM Platform 6 Smart Multicore Embedded Systems, Springer Verlag, 2013, 103-124
[2] Aracil, C. & Couroussé, D. Software Acceleration of Floating-Point Multiplication using Runtime Code Generation Proceedings of the 4th International Conference on Energy Aware Computing, 2013

Code Generation for an Application-Specific VLIW Processor with Clustered, Addressable Register Files

Research topics : Clustered VLIW, Address Generation, LLVM, Back-end compiler

I. Llopard, A. Cohen (INRIA/ENS), C. Fabre, J. Martin, H-P. Charles, C. Bernard

ABSTRACT: Modern compilers integrate recent advances in compiler construction. But most of the effort is oriented towards classical, mostly general purpose, architectures. The new constraints introduced by application-specific VLIW challenge standard strategies of code generation. We have extended a standard compiler framework with a code generation flow that tackles architecture features of VLIW processors aiming at very low-power processing.

Modern telecommunications algorithms for Software Defined Radio (SDR) demand high-performance computing and flexibility along with hard real-time requirements. Such algorithms are embedded into high-end mobile devices where low power consumption is of primary concern. To meet all these needs, a VLIW-based application-specific instruction set processor (ASIP), Mephisto, along with its code generation framework has been proposed by Bernard et al. [1]. Mephisto implements a powerful instruction set specifically designed to provide high-thoughput complex number processing. The framework has a very low level model of the architecture, directly exposing Mephisto's complexity to the programmer. In order to build an abstraction layer to ease and speed up the programming task of Mephisto, we propose a new compilation process based on LLVM, a state-of-the-art compiler infrastructure.

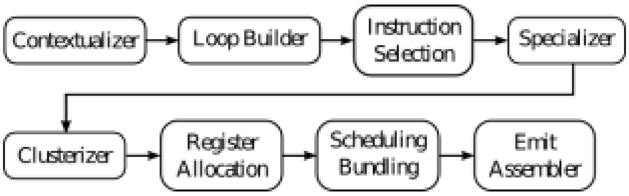


Figure 1 : General overview of the code generation flow

LLVM has already been ported to VLIW architectures. However, to our knowledge, it has not been faced to highly constrained VLIW application-specific processors. We propose new algorithms compatible with generic compilation flows like LLVM, and harnessing the partitioned memory architecture as well as clustered, addressable register files of Mephisto. Mephisto has strong partitioned register files, i.e. inter-register file communication is partial or not supported at all. It has different register banks for data and addressing purposes (pointers), each one with its associated arithmetic units: Data-specific units (MAC, Divisor, CORDICs) and Address Generators. As a consequence, data and address-related operations are two disjoint paths. It's the work of the contextualizer to separate these concerns at different instruction set abstractions (platform independent code and assembler). Based on the data-flow graph representation, the contextualizer detects conflicting paths in the graph defined by the mentioned dissociation of data and addresses computing.

Given the register file organization (indirectly addressable), we propose a new register allocation pipeline, depicted in Fig. 2. The compiler model of Mephisto does not expose to the first allocation level the fact that registers are indirectly addressed. It enables an optimal utilization of the data register file. Next allocation passes handle the real accesses to registers with an additional contextualization to isolate data from addresses as introduced early.

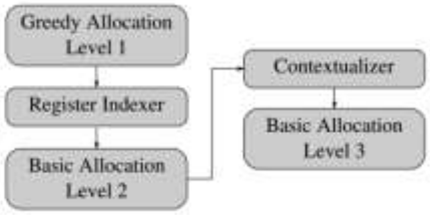


Figure 2 : Multi-phase register allocation

The main arithmetic unit is the Multiply and Accumulate (MAC) block. It's a deep pipelined operator that provides high-throughput computations on complex numbers. It introduces uncommon constraints to the compiler: not allowed copies between accumulators, non-spillable and clusterized accumulators, etc. These constraints are handled by the clusterizer which implements a dedicated clustering algorithm to manage the complexity of Mephisto's register files.

Post register allocation scheduling and bundling (instruction packets) with accurate information about instruction latencies is necessary when targeting non-interlocked processors such as Mephisto. We have also implemented a custom scheduler with additional heuristics complementing the existing VLIW-aware scheduler.

Through our compilation flow, we are able to manage the complexity of application-specific VLIW architectures such as Mephisto and automatically generate code from high-level languages.

Related Publications:
[1] C. Bernard and F. Clermidy, "A low-power VLIW Processor for 3GPP-LTE Complex Numbers Processing", Design, Automation Test in Europe Conference Exhibition (DATE), 2011.
[2] I. Llopard, A. Cohen & C. Fabre "Code Generation for an Application-Specific VLIW Processor With Clustered, Addressable Register Files", in Proc. of 10th Workshop on Optimizations for DSP and Embedded Systems (ODES'13), associated with CGO (2013). Shenzhen, China. Feb. 15th, 2013.

Hardware acceleration for Just-In-Time compilation in embedded systems

Research topics: JIT compilation, hardware acceleration, embedded systems

Alexandre Carbon, Yves Lhuillier, Henri-Pierre Charles

ABSTRACT: JIT compilation is today widely transferred in embedded systems, with significant scaling-down problems in terms of performance. In this work, we explore opportunities for a hardware acceleration of JIT compilation algorithms. Our experiments show that associative array management and dynamic memory allocation are two significant critical points in terms of performance. We propose to accelerate them with a common new ARM core functional unit. Coupled with a ARM Cortex-A5, the proposed solution achieves gains up to 25 % on the LLVM compiler and a 5x raw speedup with an area overhead under 1.4%.

To face the increasing complexity and heterogeneity of embedded systems, software stacks rely more and more on virtualization technologies, leveraging Just-In-Time (JIT) compilation. Nevertheless, the efficiency of Just-In-Time compilation depends on the ability to compensate its overhead with execution speedups of generated code.

The management of JIT compilation algorithms' complexity and irregularity on small and sparse resources (in-order processors, limited speculation, limited memory hierarchies) limits this ability and leads to important scaling-down problems in terms of performance. As a consequence, JIT compilation solutions are less attractive in this domain.

In a previous study [1], we highlighted that associative array management and dynamic memory allocation are two critical points for JIT compilation algorithms in terms of performance. In the LLVM framework compiler (LLC), they represent 25% on average of its execution time. Several software optimizations have been already proposed in the literature to address these points. The LLVM developers propose for instance 17 specialized containers, limiting the standard library usage to the less critical situations.

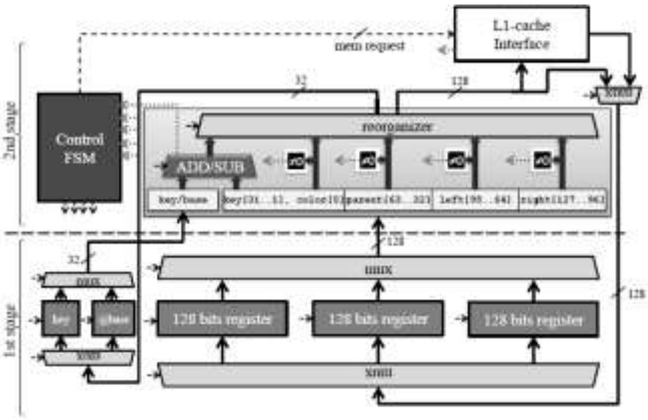


Figure 1: Block diagram of the proposed functional unit.

Based on a previous analysis of code specialization opportunities for associative array management [2], we explore hardware acceleration options for these two critical points, focusing on standard libraries to ease the solution reuse. We observe that memory allocators rely on associative arrays for their data management.

We propose a new ARM core functional unit that focuses on the STL C++ and the C's Doug Lea memory allocator with a uniform implementation of associative arrays based on Red-Black Trees [3] (Figure 1). For the performance exploration, the proposed solution is coupled with a ARM Cortex-A5 processor, implementing 9 new specialized instructions for its control.

Results on a set of benchmarks from miBench highlight gains up to 25% on LLC relative to existing software optimizations and up to 60% relative to a version only based on standard libraries. The average time spent in these two critical parts decrease from 25% to 12%. Finally, we measure a 5x raw speedup for dynamic memory allocation and associative array management with our solution, relative to their initial software implementations (Figure 2). The solution's area overhead is under 1.4% of the associated processor's area.

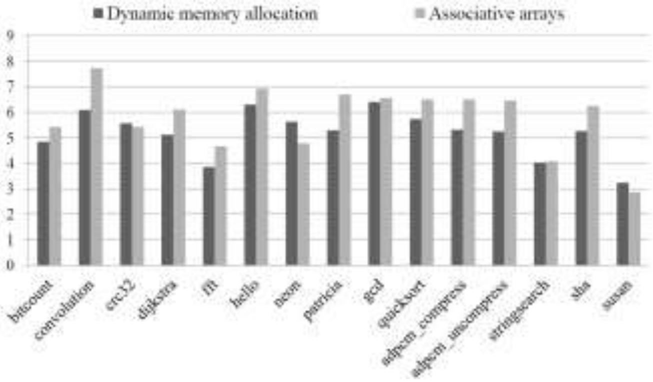


Figure 2: Raw speedup on dynamic memory allocation and associative array management with the proposed solution.

Due to the general importance of associative arrays and memory allocation in irregular algorithms, and not only JIT compilers, the proposed acceleration may also benefit to many other applications. Hidden behind standard libraries, it can be reused with a low integration effort.

Related Publications :
[1] A. Carbon, Y. Lhuillier, and H.-P. Charles, "Scaling-down to embedded systems for dynamic compilation," in 2nd International Workshop Dynamic Compilation Everywhere (DCE), 2013.
[2] A. Carbon, Y. Lhuillier, H.-P. Charles " Code Specialization For Red-Black Tree Management Algorithms", in 3rd International Workshop on Adaptive Self-tuning Computed Systems (ADAPT), 2013.
[3] A. Carbon, Y. Lhuillier, H.-P. Charles "Hardware Acceleration for Just-In-Time Compilation on Heterogeneous Embedded Systems", IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP), 2013.

PACHA: a low overhead, Platform Agnostic, Close-to-HArdware programming interface

Multi-core, Benchmarking, Performance, Energy Efficiency, Tools

Alexandre Aminot, Alexandre Guerre, Julien Peeters, Yves Lhuillier

Today, efficiently implementing an application on manycore accelerators is a hard task. *Bare metal* programming environments provided with accelerators give a low-overhead access to the platform. However, they increase the complexity of development. To reduce the programming effort on bare metal, we present PACHA. It features two aspects: a low overhead platform agnostic programming interface, which allows to handle only one version of the application code for all supported accelerators, and an easy-to-use multi-platform development environment, which abstracts the complexity of each accelerator's development environment.

Nowadays, embedded systems are exploited to execute high performance applications with strong power constraints. To meet this challenge, shared memory symmetric manycore accelerators are proposed, such as the STHORM platform created by STMicroelectronics and CEA, formerly known as Platform 2012, Tilera processors, Kalray processors, or Intel Octeon III. To exploit the platform without degrading performances, most constructors propose a close-to-hardware programming environment with similar concepts. They propose the minimum functions set to develop an application: access to memory, lock, information about the platform, traces, etc. However, these programming environments are provided by different interfaces and usages.

To abstract these differences, we propose a thin Platform Agnostic Close-to-Hardware (PACHA) programming interface [1]. The PACHA software stack is depicted on Figure 1.

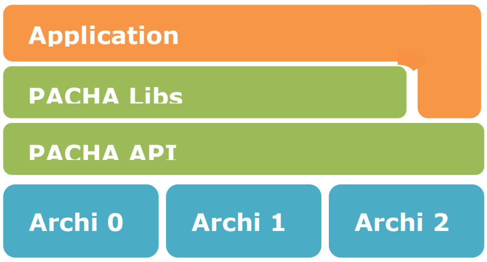


Figure 1: Software stack

The PACHA API is a low level Application Programming Interface developed in C which offers:

- A unique interface for different platforms (SMP simulators, real platforms) with a reduced API: lock, memory allocation, platform management, trace and performance
- An x86 emulator to help during the debugging phase (to check the correctness of the functionality)
- A aapid and easy way to add a new platform

The PACHA libraries consist in essential synchronization facilities and higher-level parallel programing models [2,3].

Currently 5 platforms are supported:

- x86 over Linux (pthread implementation)
- Arm Cortex A9 over Linux
- TilePro64 Linux
- TilePro64 bare-metal
- STHORM cycle-accurate simulator (with HWS support)

Figure 2 shows a usage example of what can be done with PACHA. It represents the comparison of speed/power of an application for pedestrian detection on different platforms. There is only one version of source code.

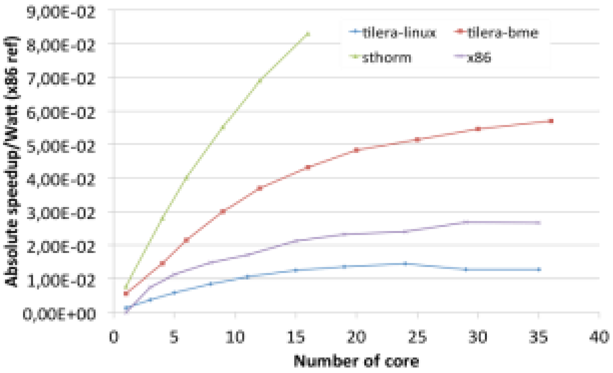


Figure 2: Demonstration on a complex vision application for pedestrian detection, comparison speed/power of multiple manycore (x86 correspond to multi-Opteron)

The PACHA environment allows multiprocessor comparison for the same target application. It offers a generic API and a unique environment for software developers. The development and the reuse of an application on bare metal environment and for multiple platforms become available at a low-cost. The overhead is under 1% on the execution. Using PACHA with the bare metal environment rather than Linux with OpenMP or Pthread, the performance gain is about 1,8x to 4x.

Related Publications:
[1] A. Aminot, A. Guerre, J. Peeters, Y. Lhuillier, "PACHA: Low Cost Bare Metal Development for Shared Memory Manycore Accelerators", *ALCHEMY Workshop, ICCS'13*
[2] M.Ojail, R.David, K.Chehida, Y.Lhuillier, L.Benini, "Synchronous reactive fine grain tasks management for homogeneous many-corearchitectures", ARCS 2011.
[3] M. Ojail, R. David, Y. Lhuillier, A. Guerre, "Artm: A lightweight fork-join framework for many-core embedded systems", DATE 2013

Early Validation of MPSoCs Thermal Mitigation through Integration of Thermal Simulation in SystemC Virtual Prototyping

Research topics : Virtual Prototyping, ESL Thermal evaluation, Thermal Management

Tanguy Sassolas, Charly Bechara, Pascal Vivet, Hela Boussetta & Luca Ferro (DOCEA Power)

ABSTRACT: The increase in power density in modern ICs leads to chip junction temperature increase and directly impacts power consumption, peak performance, ageing, and design costs. For MPSoC architectures, thermal profiles are dependent on the application case, the data set and the scheduling. Thus, thermal phenomena are hard to predict, yet control. In this paper we present our virtual prototyping environment for the early evaluation and mitigation of thermal phenomena in MPSoC. Our solution is based on an efficient co-simulation between a functional SystemC based simulator and a thermal evaluation tool developed by our partner DOCEA Power.

As part of their 3 years-long collaboration, DOCEA Power and the CEA jointly developed a co-simulation technology between the CEA virtual prototyping platforms, SESAM and TGV, and DOCEA Power's power and thermal evaluation tools. This strong partnership allowed us to demonstrate during DAC'13 conference this key technology on DOCEA Power's booth (Fig.1). The demonstration comprised the evaluation and thermal mitigation of a quad core architecture executing a pedestrian detection application. This innovative demonstration was coupled with a technical presentation during the conference industrial tracks [1].

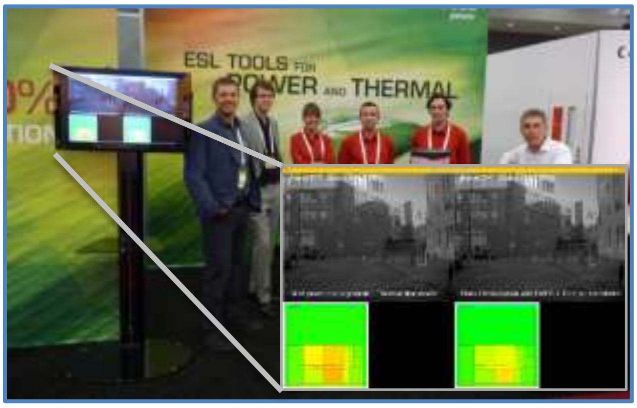


Figure 1: Thermal mitigation demonstration developed by the CEA on DOCEA Power's DAC'13 booth

To allow thermal mitigation development without increasing the time to market, thermal solutions must be seeked at the Electronic system level. Such solutions are of course to be refined with the design. An efficient ESL thermal tool is needed to allow such thermal mitigation development and it shall provide links with power and functional ESL tools as these three domains are heavily linked. Indeed the temperature has an impact on power consumption and on the mitigation scheme i.e. the applicative scenario. This applicative scenario has in turn an impact on the power consumption; the power consumption impacting thermal evolution.

Thermal design tools come in numbers, however most industrial tools such as Flowtherm, SolidWorks or ANSYS fluence, use detailed fluid dynamics simulation. Although they are well-suited for packaging design as extremely accurate, they lack the speed and interfaces to use them for platform level exploration. Some ESL tools do exists and have been widely used in the academic world among which HotSpot (Univ. of Virginia) and 3D-ICE (EPFL) however they are designed for single chip dissipation studies, and use complex thermal grid which hamper transient simulation speed. But most of all, these tools did not natively support the modeling of the impact of temperature on power and application. The closing of the simulation loop between power/thermal/ and functional was missing.

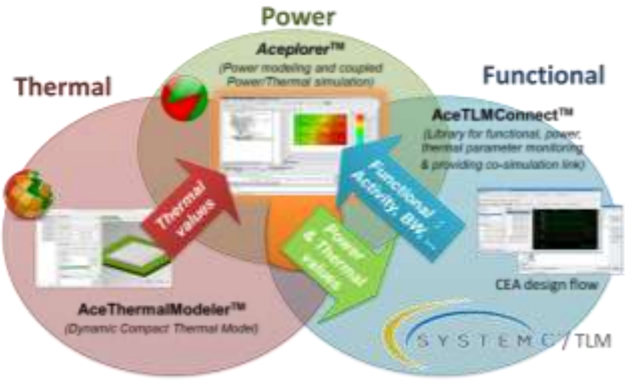

Power
Aceplorer™
(Power modeling and coupled Power/Thermal simulation)
Thermal
Functional
AceTLMConnect™
(Library for functional, power, thermal parameter monitoring & providing co-simulation link)
Thermal values
Functional Activity, BW, ...
Power & Thermal values
CEA design flow
AceThermalModeler™
(Dynamic Compact Thermal Model)
SYSTEMC/TLM

Figure 2: ESL Modeling & Simulation Framework

Our thermal evaluation flow is depicted on Figure 2. A dynamic compact thermal model based on the geometry and thermal properties of the package and die is generated with AceThermalModeler. The power modeling is performed with Aceplorer which also enables coupled power and thermal simulations. AceTLMConnect Library is used to monitor functional parameters in the CEA simulation platform, such as bandwidth and activities, in order to stimulate the power and thermal models.

Using this flow we studied the behavior of a pedestrian detection application running on a quad XP70 architecture developed by the CEA. We confronted various mitigation schemes to select the one that guaranteed both the respect of the circuit's thermal envelope and the applications quality of service. The overall execution time was limited to 21 minutes for a 9 minutes simulated scenario thanks to a PVT virtual prototype designed specifically for thermal phenomena time scale [2] and based on HARS[3] runtime.

Related Publications :
[1] Presentation available at DAC conference archives: http://www.dac.com/App_Content/files/50/50th%20DT%20Slides/06D_1.pptx
[2] T. Sassolas, C. Sandionigi, A. Guerre, A. Aminot, P. Vivet, H. Boussetta, Luca Ferro and Nicolas Peltier, "Early Design Stage Thermal Evaluation and Mitigation: The Locomotiv Architectural Case", Design, Automation and Test in Europe (DATE), 2014.
[3] Y. Lhuillier et al., "HARS: A hardware-assisted runtime software for embedded many-core architectures," ACM Transactions on Embedded Computing Systems, to be published.

A Fast and Accurate Methodology for Power Estimation and Reduction of Programmable Architectures

Research topics : Low-power, RTL design, processor, design space exploration

Erwan Piriou, Raphaël David and Fahim Rahim, Solaiman Rahim (Atrenta)

ABSTRACT: We present a power optimization methodology that provides a fast and accurate power model for programmable architectures. The approach is based on a new tool that estimates power consumption from a register transfer level (RTL) module description, activity files and technology library. It efficiently provides an instruction-level accurate power model and allows design space exploration for the register file. We demonstrate a 19% improvement for a standard RISC processor.

An early power estimation and exploration tool is a strong asset for optimizing architecture. Taking into account the emergence of specialized and/or general purpose microprocessor resources in recent system-on-chip (SoC) designs, the need for fast power estimation becomes crucial. In particular, considering embedded processors, the opportunity to characterize an instruction set architecture (ISA) allows the feedback of important power figures to the compiler designer. Additionally, from a hardware point of view, the capability to highlight the cost/impact of architectural choices (e.g., the register file architecture) enables analysis of the architecture in terms of power consumption.

The proposed approach considers the instruction set architecture and the RTL description as entry points. It allows the acquisition of a power scorecard for the programmable core at the instruction level while considering the architecture description at the RT level. In fact, several power scorecards could be generated depending on the chosen technology and frequency.

This approach proposes to drastically reduce the computing time to rapidly obtain a power model and architectural hints for the programmable core design while remaining accurate.

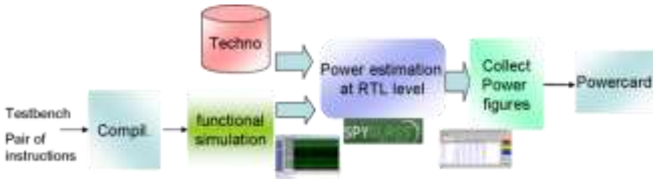


Figure 1 :Power estimation flow for programmable architectures

First, testbenches for all possible legal instruction pairs are generated in C code using asm volatile pragma features. Then, using the compilation toolchain, the program is compiled and the executable memory footprints are generated for RTL simulation. Next, simulations are performed for the hardware architecture using an RTL simulator. The activity files (.vcd, .fsdb or .saif format) are generated to feed the Atrenta SpyGlass Power Estimate tool. Finally, all RTL and activity files are processed by the Atrenta SpyGlass RTL Power Estimation and Reduction tools. Concurrently, the technology library is linked and a functional frequency is initialized. The tool outputs the power figures for all testbenches. They are gathered to build a power scorecard. From a power analysis, the microarchitecture power budget is analyzed. The register file constitutes the greediest module.

16/32bit ISA of the AntX processor [1] is considered as a case-study. For a core handling at least 74 instructions, 74 benches are required for intrinsic power characterization and 2701 benches for pair of instructions estimations. Therefore, it is not realistic to quickly provide a complete gate-level power model for the processor. In fact, an aggregate testbench would have a length of 1 millisecond for the activity file. A single analysis at the netlist level of an activity file of such a length cannot be planned.

In this case study, the developed power estimation flow at the RT level allows a speedup by a factor 15 compared to gate level methods for the RISC processor running at 400 MHz on a TSMC 45nm library. The powercard contains all average power consumption obtained for the execution of the same instructions and the execution of all possible pairs of instructions. They are classified by functionality, by format and by data access. These figures have been used to back annotate the standalone instruction set simulator of the chosen processor.

An average accuracy of nearly 80% was obtained for the power model on a representative set of benchmarks executing motion estimation, discrete wavelet transform, and a sort algorithm compared to gate-level characterization. Experiments have also been performed for unified/split register file architectures and different numbers of registers (physical/compiling option).
Considering the unified version, when the number of used registers is shifted from 16 to 8, the total design and the register file power consumption remain the same. However, the execution time increases by 5.6. The splitting and the clock gating of one part of the RF can save 19% of the power consumption.

This opportunity for early design tools to consider programmable architectures power consumption is a promising solution to make the design process easier. It is of particular importance for compiler and hardware designers when pursuing power saving improvements. In the future, the approach would be applied on the other part of the microarchitecture.

Related Publications:
[1] Charly Bechara et al. A small footprint interleaved multithreaded processor for embedded systems. In ICECS, pages 685–690, 2011.

Homomorphic Encryption in Cloud Computing

Research topics : homomorphic encryption, cloud computing, crypto computing

S. Fau, R. Sirdey, S. Carpov, C. Aguilar-Melchor (XLIM), C. Fontaine (Lab-STICC/CNRS/Télécom Bretagne), G. Gogniat (Lab-STICC)

ABSTRACT: Providing security in the cloud has been approached in different angles and with different upsides and downsides in the trade-off between security and efficiency. One of the most serious candidate is Homomorphic Encryption, as it allows to make some computations on the encrypted data, while its security has been proven to be very strong.
Our work consists in identifying the computational hot spots resulting of the use of Homomorphic Encryption and find ways to mitigate them. Eventually we demonstrated that Homomorphic Encryption was already a realistic approach (and probably the best) for some algorithms.

Homomorphic encryption has been introduced in 2009 and it has been one of the most promising cryptographic research line since. Indeed, specific computations on the ciphertexts can be passed on to the underlying plaintexts thanks to the special design of a homomorphic encryption scheme.
For example, we can see in Fig.1 the main advantage of homomorphic encryption compared to classical encryption. When the ciphertexts cannot be used for anything else than decryption in classical encryption, they can be added or multiplied when using homomorphic encryption, and the additions or multiplications impact on the plaintexts.

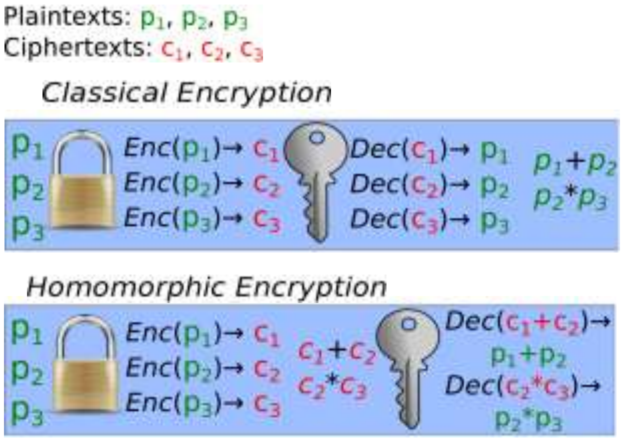

Plaintexts: p_1, p_2, p_3
Ciphertexts: c_1, c_2, c_3
Classical Encryption
p_1 p_2 p_3
$Enc(p_1) \rightarrow c_1$
$Enc(p_2) \rightarrow c_2$
$Enc(p_3) \rightarrow c_3$
$Dec(c_1) \rightarrow p_1$
$Dec(c_2) \rightarrow p_2$
$Dec(c_3) \rightarrow p_3$
p_1+p_2
p_2*p_3
Homomorphic Encryption
p_1 p_2 p_3
$Enc(p_1) \rightarrow c_1$
$Enc(p_2) \rightarrow c_2$
$Enc(p_3) \rightarrow c_3$
c_1+c_2
c_2*c_3
$Dec(c_1+c_2) \rightarrow p_1+p_2$
$Dec(c_2*c_3) \rightarrow p_2*p_3$

Figure 1 : Classical encryption vs homomorphic encryption

This property is particularly interesting in the cloud computing paradigm. In the client/server model, a client encrypting its data using homomorphic encryption allows a server to perform computations on said data, without gaining any knowledge on it. Homomorphic encryption thus appears to be the perfect solution to guarantee outsourced data security, while keeping the possibility to process it.
Of course, using homomorphic encryption efficiently in the cloud computing models requires a lot of work on both the cryptographic primitives themselves and on the process of homomorphic-encrypted data. Our research focuses on the latest issue, and has been published in two articles in 2013 [1,2].

Our first concern was to gauge the possible computations that could be done on homomorphic-encrypted data. Starting with a homomorphic encryption scheme that allowed XOR and AND operations on encrypted bits, we wrote a small set of simple algorithms, using only XOR and AND gates, that can now be run in the encrypted domain, such as a bubble sort, a threshold, a FFT, etc. [1].
We built our platform in order to be relatively independent of the homomorphic encryption scheme used, since the state-of-the-art is not stable yet. Indeed, we can use a new external implementation of a homomorphic encryption scheme without changing the non-cryptographic part of our platform.
Once we had seen the extent of the expressivity we could achieve using homomorphic encryption, we focused on reducing the overhead of homomorphic computation on the non-cryptographic end.
All the computationally realistic homomorphic encryption schemes are based on lattice-based cryptography, and have the same computational hot spots.
When computing on homomorphic-encrypted data, the XOR gate is nearly free in comparison to the AND gate, which has a big overhead. Another limiting factor is the multiplicative depth of the algorithm we want to run in the encrypted domain (by multiplicative depth we mean, seeing an algorithm as a boolean circuit, the maximum number of AND gates in a path, for all the paths).
We showed in our most recent article [2] how the number of AND gates and the multiplicative depth are crucial to the efficiency of computation on homomorphic-encrypted data. Minimizing theses values are therefore a very big perspective for our future research.
We already started to rebuild Boolean circuits in order to have less AND gates, and smaller multiplicative depth on some test algorithms. In this manner, we have been able to demonstrate a (fictitious) medical diagnosis algorithm that compares (encrypted) medical test results from a patient (blood pressure, cholesterol, etc.) to standard values and give a risk/no-risk answer. Using homomorphic encryption, such a test was completed in less than 2 seconds on an 8-cores processor with a security equivalent to 128-bits.
We have strong hope to be able to perform in the close future more sophisticated algorithms in less time, while keeping the security to the same level.

Related Publications :

[1] Aguilar-Melchor, C.; Fau, S.; Sirdey, R.; Fontaine, C. & Gogniat, G. (2013), 'Recent advances in homomorphic encryption: a possible future for signal processing in the encrypted domain', IEEE Signal Processing Magazine 30(2), 108-117.
[2] Fau, S.; Sirdey, R.; Fontaine, C.; Aguilar-Melchor, C. & Gogniat, G. (2013), 'Towards Practical Program Execution over Fully Homomorphic Encryption Schemes''Proceedings of 3PGCIC 2013'.

3

Power & Temperature Optimized Digital Circuits

DVFS control

FDSOI UWVR circuits

Low-Power Architectures

Fine-Grained Adaptive Local Voltage Dithering in a 32 nm GALS MPSoC

Research topics : GALS, PVT monitoring, AVFS, process compensation, multi-core

E. Beigne, I. Miro-Panades

ABSTRACT: With deep submicron technologies, circuits are suffering from large variations impacting yield and the overall energy efficiency. In particular, within-die PVT variations have a strong impact on complex MPSoC as neighboring processors can have disparate maximum frequencies and/or power. In this work, a 32 nm CMOS MPSoC circuit is proposed to demonstrate a fine-grained dynamic adaptation to PVT variations to track an optimum Voltage and Frequency (V/F) functional point in each Processing Element (PE). It embeds an innovative AVFS approach based on fast frequency re-programming and voltage hopping.

Contrary to classical AVFS architectures where frequency is fixed and voltage is scaled, in this proposal, voltage supplies (VH, VM, VL) are fixed and frequency is dynamically adapted to PVT variations at each voltage point. Each PE is thus able to reach its optimal energetic V/F point. Finally, to reduce the energy per instruction, voltage is efficiently dithered between the three adapted V/F points.

LoCoMoTIV chip is composed of 4 single-issue in-order STxP70 PEs, a DMA engine, a 256KB shared SRAM memory, a Hardware Synchronizer (HWS), an Asynchronous NoC (ANoC), and an AXI bridge. Each module is an independent Globally-Asynchronous and Locally-Synchronous (GALS) clock and power domain allowing the local fine-grained adaptation. In this implementation in ST CMOS032LP, two PEs are instrumented with the full adaptive scheme previously described for high energy efficiency and can run up to the maximum speed of 947MHz in typical conditions.

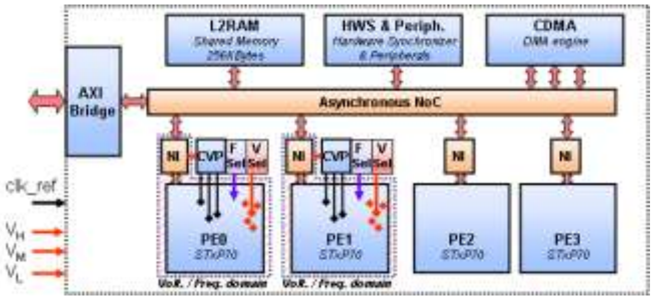


Figure 1 : LoCoMoTiV GALS MPSoC adaptive architecture

The fine-grained proposal requires low-cost local sensors, local V/F actuators and a local control loop for adaptation. Local dynamic PVT diagnosis is done with Timing Fault sensors (TMFLT) placed on critical paths of the processor, coupled to digital ring oscillators (MPROB). Typically, TMFLT are used first during test-at-speed for calibration to determine maximum PE speed for the three supply voltages. Then, during operation, those frequencies are adjusted by a fast Frequency Locked Loop (FLL) according to sensors information. A Clock/Variability/Power Controller (CVP) is in charge of V/F point selection and voltage-hopping duty-ratio control according to an applicative Operating Point (OPtarget). TMFLT objective is to prevent from an error by raising an interrupt when setup slack time comes near zero.

Small (10.68 µm²) shadow latches are directly connected to D-inputs of flip-flops and detect a transition of the signal in the window directly implemented in the clock tree.

Besides, MPROB is a low area (450µm²) distributed macro-block embedding several Ring Oscillators (ROs) made of inverters, long wires, latches, XOR gates, large capacitive nets and a temperature dependent RO based on current-starved inverters biased by a thermally dependent current generator. The resulting voltage and temperature cartographies can then be processed using statistical test hypothesis algorithms.

The FLL is a 3300um² all-digital clock generator able to generate Fcore frequency from 2 GHz down to 15 KHz using a reference frequency of 100MHz. Full frequency point reprogramming takes less than 180ns without interruption of the functional clock. Vdd-hopping circuits are distributed in each PE to generate internal Vcore supply voltage with low IR drop. Each Power Hopping-Switch (HS) block delivers 10mA and takes 1024µm². The CVP (~13000µm²) plays the role of adaptation controller and power sequencer. It collects all the diagnosis signals and adjusts voltage and frequency to minimize power consumption, while respecting applicative constraints.

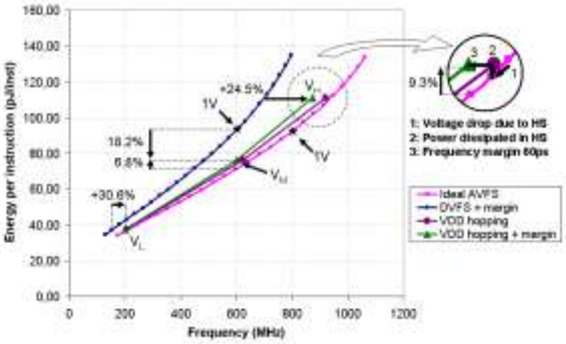


Figure 2 : Energy gains results

In typical PVT conditions, the proposed Vdd-hopping implementation saves up to 18% in energy per instruction compared to the classical worst-case design using 25% timing margins for safe operation (Figure above).

The LoCoMoTIV chip applies the AVFS locally, which allows adapting to each processor needs, and saving as high as 34% in case of unbalanced loads between two processors. Similarly, considering in-die process variability, decoupled Vdd-hopping also allows to save 7.2% energy/operation for a 10% difference in PE timings. PE adaptation instrumentation for local adaptation costs around 10% of the total PE in terms of area.

Related Publications:

[1] Beigne, E.; Miro-Panades, I.; Thonnart, Y.; Alacoque, L.; Vivet, P.; Lesecq, S.; Puschini, D.; Thabet, F.; Tain, B.; Benchehida, K.; Engels, S.; Wilson, R.; Fuin, D., "A fine grain variation-aware dynamic Vdd-hopping AVFS architecture on a 32nm GALS MPSoC," ESSCIRC 2013, pp.57-60, 16-20 Sept. 2013

Coupled Voltage and Frequency Control for DVFS Management

Research topics: Power Management, DVFS, advanced control techniques

M. Altieri, W. Lombardi, D. Puschini, S. Lesecq

ABSTRACT: During the last decade, Dynamic Voltage-Frequency Scaling (DVFS) techniques have been widely proposed to improve integrated circuit efficiency. When they are based on coupled drivers, the actuators need to be jointly designed. In the present work, a control mechanism is proposed to jointly control the voltage and frequency transient periods when both actuators are developed independently. Implemented in STMicroelectronics 32nm bulk and 28nm FDSOI technologies, it requires a small silicon area and power consumption overhead. Experimental results with two independently developed drivers are provided.

Power consumption is a limiting factor in Very-Large-Scale Integration (VLSI) circuits, especially for mobile applications. In the last decade, several low-power design techniques have been proposed. Dynamic Voltage and Frequency Scaling (DVFS) has proven to be highly effective to reduce the power consumption of embedded systems while fulfilling the performance requirements. Basically, DVFS is composed of a variable voltage engine and a variable clock generator. These drivers need to be dynamically controlled to reduce the power consumption while maintaining the performance required by the application running on the computing platform. Basically, the clock frequency should be set considering the supply voltage value or vice versa. For a given voltage, a high frequency produces timing faults in the circuit logic while a low frequency will result in poor calculation performance with high power consumption. For these reasons, the management policy of both actuators [1] must ensure a predefined sequence to switch from one voltage-frequency state to another one. For instance, for an increasing frequency step the management policy must firstly increase the supply voltage, and once its output is stabilized, the frequency can be increased. The timing of this sequence depends on the dynamic response of both actuators.

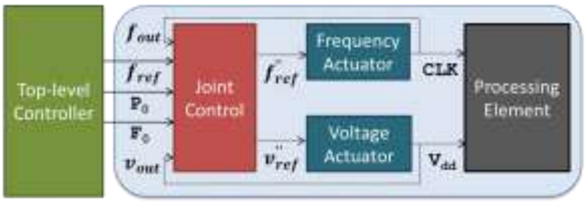


Figure 1 : DVFS system with joint control

The present work [1] deals with a coupled control of the voltage and frequency drivers (see Figure 1). This control does not require the implementation of temporal sequences as used when both drivers are designed independently. Moreover, the joint control proposed here is able to manage actuators with different dynamics. Requiring only a linear function definition and a target frequency, it makes both actuators evolve simultaneously in a robust and energy-efficient way.

The proposed control has been implemented in STMicroelectronics 32nm bulk 28nm FD-SOI technologies using standard cell methodology and validated through mixed-signal simulations. Figure 2 shows the evolution of the clock frequency F and supply voltage V, respectively without (green) and with (orange) the coupled control. The functionality limit is given by the purple curve. As expected, the coupled control gets the chip functions closer to the functionality limit.

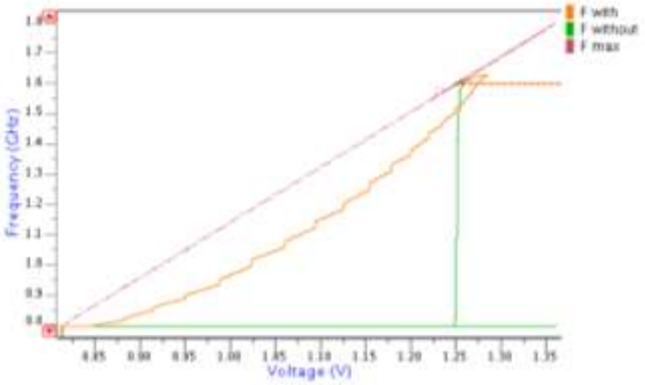


Figure 2: Joint Control Architecture

The synthesis report, presented in Table 1, shows that this implementation requires a relatively small area and is power efficient, compared to the 32nm bulk implementation.

Technology		32nm bulk	28nm FD-SOI
Clock Frequency (MHz)		100	
Critical Path Length (ns)		9.76	5.62
Number of Cells		1556	1039
Area (μm^2)	Combinational	1048.2336	583.9296
	Noncombinational	301.2672	193.4464
	Total	**1349.5008**	**777.376**
Power (μW)	Internal	34.5	21.0
	Switching	28.3	8.02
	Leakage	0.05	2.23
	Total	**62.9**	**31.2**

Table 1 : Synthesis report

Related Publications :
[1] Mauricio Altieri, Warody Lombardi, Diego Puschini, Suzanne Lesecq, "Coupled voltage and frequency control for DVFS management", 3rd International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS), Karlsruhe, Germany, September 9-11, 2013, 2013.

Thermal-Aware event-based DVFS control

Research topics : event-based control, DVFS, thermal

S. Lesecq, E. Beigné, D. Puschini, S. Durand (GIPSA-Lab)

ABSTRACT: Dynamic Voltage and Frequency Scaling techniques are widely used in order to decrease the power consumption of a platform while meeting the required performances. Unfortunately, they do not usually take into account the thermal aspects that have significant effects on the power consumption of circuits manufactured in advanced technologies. In order to mitigate the thermal effects, a new DVFS scheme, applied in a chopped way, is proposed. This scheme is extended to an event-based one in order to reduce the control and communication burdens. Simulation results are promising and evaluation on a Silicon platform must be carried on .

The upcoming generations of embedded integrated systems have reached limits in terms of power consumption, computational efficiency and fabrication yield. One of the problems induced by advanced technologies (32nm and beyond) is the so-called process variability: the chip performance cannot be ensured from one die to another, nor over a single chip. More generally, on chip Process, Voltage and Temperature (PVT) variations, unbalanced workload and ageing evolution yield to different spatial and temporal dynamics. Moreover, clock distribution has become highly difficult, especially for Many Processor Systems on Chip (MPSoCs) for which high speed analog clock signals should be routed around the system, the clock tree suffering from signal integrity issues and layout difficulties, leading to an increase in the design cost and a decrease in the yield. Globally Asynchronous Locally Synchronous (GALS) architectures alleviate this clock distribution difficulty, the chip being split into several frequency islands. Moreover, they allow each synchronous island to be supplied with a different voltage. As a consequence GALS architectures are suitable for fine grain power management as the power consumption of the whole platform depends on the supply voltage and the clock frequency applied to each VFI.

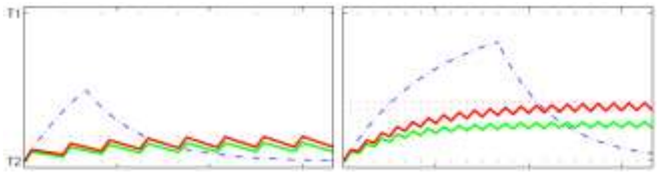


Figure 1 : DVFS control schemes (blue: classical DVFS scheme, red: non-linear thermal-aware DVFS, green: linear thermal-aware DVFS)

Dynamic Voltage and Frequency Scaling (DVFS) methods are designed in order to provide just enough power to a given VFI for executing the tasks that run on it in such a way that they meet their deadline. This allows guaranteeing the overall performance while minimizing the power consumption. Closed-loop feedback control techniques can hence be efficiently implemented to achieve such energy-performance tradeoff.

However, for upcoming mobile platforms designed in advanced technologies, the temperature variations must be as well controlled, or at least limited. Actually, the leakage power, which is a significant contributor to the total power consumption, highly depends on the temperature. Consequently, thermal effects have to be taken into account in the DVFS control law. We developed a nonlinear thermal-aware DVFS strategy based on a chopped-version of the classical DVFS schemes in order to mitigate the thermal effects [1]. The temperature asymmetric dynamic behavior has been taken into account in the modeling phase. Fig. 1 shows simulation results when a classical DVFS scheme is applied and when the thermal-aware DVFS strategies are implemented (resp. with symmetric and asymmetric thermal dynamics behaviors).

Note that the thermal-aware control scheme has been extended to an event-based (asynchronous) approach [2], leading to a decrease in the control burden.

We also proposed a control scheme of the energy-performance tradeoff for multiprocessor systems under strong variability constraints [2]. It implements an event-based technique (see fig. 2) on top of a robust DVFS approach. Thus, the control law is computed and updated only when a certain ratio of the workload to treat is reached. As a consequence, the periodicity of computations for the same final performance is relaxed. Furthermore, in the case the control loop is closed over a communication link (Networked Control System - NCS), it reduces the communication exchanges.

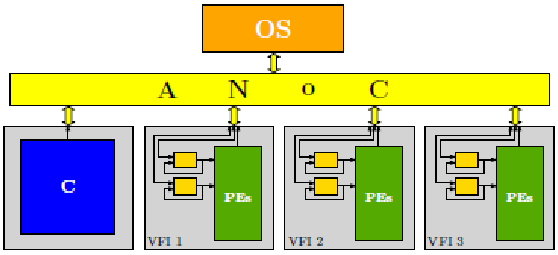


Figure 2 : Event-based architecture principle

Related Publications:
[1] S. Durand, S. Lesecq, " Nonlinear and Asymmetric Thermal-aware DVFS Control", European control Conference, Zurich, Switzerland, 2013.
[2] S. Durand, S. Lesecq, "Asynchronous Thermal-aware DVFS Control", American Control Conference (ACC), Washington, USA, 2013.
[3] S. Durand, S. Lesecq, E. Beigné, D. Puschini, "Event-based DVFS Control in GALS-ANoC MPSoCs", American Control Conference (ACC), Washington, USA, 2013.
[3

Power and thermal modeling of many-core systems, online strategies for temperature and energy consumption minimization

Research topics : Many-core systems, power & thermal model, online strategies

T. Ducroux(ST), L. Vincent, M. Becher, S. Bensalem (UJF), F. Pacull, J. Mottin, S. Lesecq

ABSTRACT: Power consumption and heat dissipation are crucial concerns in advanced modern embedded many-core systems. Because of their intimate interdependency, these two phenomena have to be studied altogether at both hardware level and software level. A novel approach is proposed to provide the application with on-line thermal estimation based on complex sensor fusion, a comprehensive power model of full many-core system is described and finally a framework named Icy-Core utilizing the previous assets is proposed to master heat problems in many-core system using thermal mitigation techniques.

One challenging issue in many-core systems is to master the effects of variability at a fine grain resolution, to be able to leverage maximum potential of the platform. To do so, low-cost sensors have been designed using digital structures such as ring oscillators whose frequencies depend on the operating point of the circuit. Accessing the temperature value using such sensors is however not straightforward and require additional post-processing. We present a method that allows temperature estimation from low-cost digital sensors [2]. This method relies on Kolmogorov-Smirnov goodness-of-fit hypothesis test: after a fusion phase, frequency values are compared to a data base of reference models as shown on Fig. 1. This method achieves an average absolute error of 6.2 °C which is comparable with state-of-art high cost absolute thermal sensors.

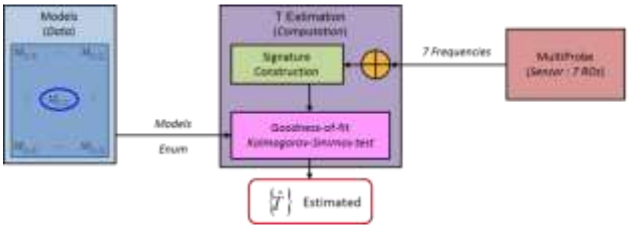


Figure 1: Temperature Estimation Technique

Another challenging issue while programming embedded many-core systems is to precisely monitor the impact of an applicative scenario on its power budget. To simulate complex applications running over many-core systems, fast back-annoted simulators have been proposed using instruction level power analysis [1]. The power model we propose has been calibrated against low-level RTL simulation environment, using Prime Time PX power analyzer. The overall error of the entire many-core system in power estimation is below 7.5%, and for the error in the power model of the processor is below 3.8%.

The Icy-Core framework [3] allows to monitor and test task migration strategies, according to the thermal behavior of a running program.

Differently from the existing simulation frameworks, the purpose of this framework is to ease the analysis of a dynamic reconfiguration solution in order to face thermal issues. As shown in Fig. 2, the framework is decomposed on three phases (initialization, main loop and visualization) and five modules. There are two existing modules: a simulator of targeted many-core platform called STHORM (formerly P2012), and a thermal simulator (3D-ICE). There are also two specially built modules, a power profile generator, a heat map visualization module, and the task migration algorithm which we want to assess.

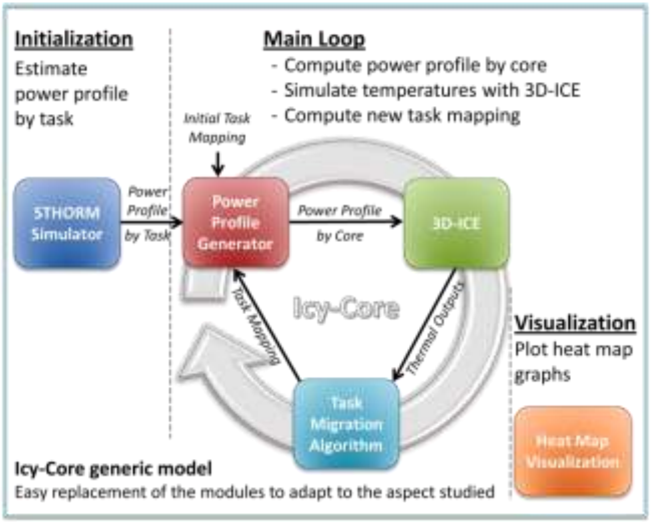


Figure 2: The Icy-Core Framework

The advantage of the Icy-Core architecture is its generic model. Thus, it is used to test different algorithms against the same constraints like the thermal stress, helping us evaluating their reaction. Moreover, it can be used to evaluate other dynamic reconfiguration strategies against other constraints than thermal. For instance, for resource management purposes such as memory management, the thermal simulator can be replaced by other module that computes memory utilization.

Related Publications:
[1] Ducroux, T.; Haugou, G.; Risson, V. & Vivet, P. (2013), 'Fast and accurate power annotated simulation: Application to a many-core architecture''Proceedings of the Power and Timing Modeling, Optimization and Simulation (PATMOS), 2013
[2] Vincent, L.; Maurine, P.; Beigne, E.; Lesecq, S. & Mottin, J. (2013), 'Temperature and Fast Voltage On-Chip Monitoring using Low-Cost Digital Sensors', 4th European Workshop on CMOS Variability (VARI 2013), Karlsruhe, Allemagne.
[3] Becher, M.; Bensalem, S. & Pacull, F. (2013), 'Icy-Core environment for simulating thermal effects of task migration algorithms on multi- and many-core architectures', Proceedings of the Platform 2012/STHORM embedded many-core acceleration (DATE-2013)

Analog Encoded Neural Network for Power Management in MPSoC

Research topics : power management, MPSoC, DVFS, neural networks

B. Larras*, B. Boguslawski, C. Lahuec*, M. Arzel*, F. Seguin*, F. Heitzmann, (*Telecom Bretagne)

ABSTRACT: Encoded neural networks mix the principles of associative memories and error-correcting decoders. This work introduces an analog implementation of this new type of network to manage the power distribution in Multiprocessor System-on-Chip (MPSoC). The proposed circuit has been designed for the 1V supply ST CMOS 65nm process, with a low complexity and low power consumption. Compared to a digital counterpart based on game theory, this analog solution consumes 6800 times less energy and reacts 4500 times faster. Thus, allows to fully exploiting DVFS circuits switching capabilities to adapt the power distribution of an MPSoC.

Multiprocessor Systems-on-Chip (MPSoCs) gained a lot of importance in recent years. Thanks to their distributed and scalable architecture they offer high performance required in real-time applications with potential power savings allowing fulfilling energy restrictions under battery operation. An 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 Vdd 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 (Vdd,f), or power modes, based on workload, latency, temperature, … A dynamic approach such as game theory has been used to control PEs frequencies at runtime [1]. However, the time response of the control unit provides a solution for (Vdd,f) from few µs to hundreds of µs [2]. With the growing speed of electronic devices, this implies that the decision on (Vdd,f) is obsolete when taken. Moreover, power management techniques have been greatly improved potentially allowing switching between two different frequency/voltage power modes in time of the order of tens of nanoseconds. Consequently, the time response of the control unit limits exploiting fast reaction possibilities of DVFS circuits.

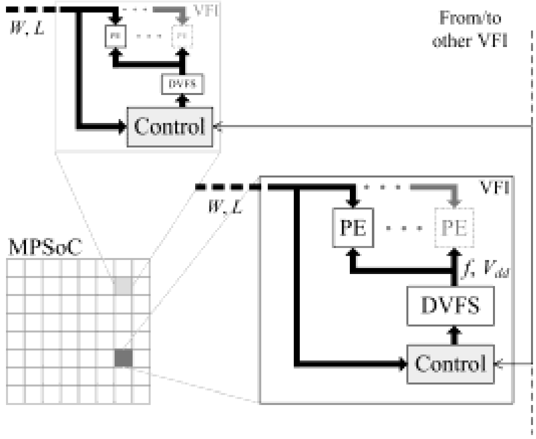


Figure 1 : MPSoC with power management capability; W – workload, L – latency, f – VFI clock frequency, Vdd – VFI supply voltage.

Recently Gripon and Berrou proposed a new model of networks (encoded neural networks - ENN) that are able to learn and recall information in the presence of noise, or to find the closest known solution in case the information is partially unknown. This work proposes to use analog ENN for power management in MPSoCs. The control unit in Fig. 1 is replaced with a part of an ENN and is fed with the parameters regarding the working conditions as workload W and latency L. Then, the ENN converges and assigns an adapted clock frequency f (through a DVFS circuit) for each VFI. The network is designed for the ST CMOS 65nm design kit. The supply voltage is 1V. The circuit needs 10.25pJ per VFI to associate a clock frequency to working conditions. The goal – computing clock frequencies for an MPSoC in tens of nanoseconds – is reached at the expense of area increase of the control unit. This is acceptable considering all the huge gains in terms of speed and energy consumption (×4500 and ÷6800 respectively compared to game theory) this solution offers, Table I.

Table 1:. Comparison with state-of-the-art

	[2]	This work
Number of VFIs	64	64
Surface (mm², per VFI)	0.014	0.092
Time response (µs)	120.32	0.027
Energy consumption per decision (nJ, per VFI)	68.6	0.01
Technology	65nm	65nm

This work shows that the time response of a simulated analog ENN matches the requirements to exploit the fast reactivity of DVFS circuits. Analog ENN also offers low power and acceptable complexity.

Related Publications:
[1] D. Puschini, F. Clermidy, P. Benoit, G. Sassatelli, L. Torres, "Dynamic and Distributed Frequency Assignement for Energy and Latency Constrained MP-SoC", Design, Automation & Test in Europe Conference & Exhibition (DATE), 2009.
[2] I. Mansouri, C. Jalier, F. Clermidy, P. Benoit, L. Torres, "Implementation Analysis of a Dynamic Energy Management Approach Inspired by Game-Theory", IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2010.
[3] B. Larras, , B. Boguslawski, C. Lahuec, M. Arzel, F. Seguin, F. Heitzmann, "Analog encoded neural network for power management in MPSoC", 2013 IEEE 11th International New Circuits and Systems Conference, NEWCAS 2013.

Methodology for Power Mode selection in FD-SOI circuits with DVFS and Dynamic Body Biasing

Research topics: Power Management, DVFS, Body biasing

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

ABSTRACT: Embedded systems need ever increasing computational performances. Since they have limited energy resources, power consumption has to be minimized. Dynamic Voltage and Frequency Scaling (DVFS) techniques combined with Body Biasing decrease the power consumption of a chip by allowing just enough chip performance so that tasks meet their deadline. Executing tasks with the two frequencies neighbor of the target frequency is supposed to minimize the power consumption. Unfortunately, this choice is not always optimal when 3 actuators are considered, with at least one of them with a discrete set of values available.

The ever increasing need in computational performance for embedded systems implies an increase in the clock frequency applied to the chip. BULK transistors' downsizing has allowed integrating more features while increasing the computational performance. However, nowadays, the physical limits of downsizing have been reached for bulk technology. As a consequence, new technologies (e.g. FD-SOI, FinFET) have been developed to cope with the downsizing and performance issues. These new technologies provide a way to act on the threshold voltage of the Processing Element (PE) by modifying the body bias voltage (Vbb). Therefore, a new "actuator" is added in FD-SOI technology to manage the power-performance trade-off.

Here, the main issue is to minimize the power consumption under performance constraint when 3 actuators are considered to feed the PE. Discrete actuators are considered for the supply voltage (Vdd), the clock frequency actuator (F) and the body bias voltage (Vbb). Therefore a small set of values is available for the voltages and the clock frequency. The notion of Power Mode (PM) has been introduced to make the methodology independent from the actuators implemented. A PM is defined as the couple (F, P) where F is the clock frequency and P is the total power consumption associated to F. The frequency can be achieved by adjusting Vdd and/or Vbb.

It has been proved that the power consumption is minimized by executing a given task with the 2 neighbor frequencies of the target frequency (Ftarget) if the curve P vs. F (denoted hereafter P(F)) is convex. However, P(F) does not always fulfill the convexity property when 3 actuators are combined.

The method we developed is made of 2 steps. The first one consists in selecting the PMs which belong to the Discretely Convex Subset (DCS). The second step is the task execution, where a task will be executed with the PMs belonging to the DCS. Different situations may appear when executing a task:

- if Ftarget is available on the PE and if the PM corresponding to Ftarget belongs to the DCS, the PM is applied directly in order to execute the task;
- if Ftarget is available on the PE and if the PM corresponding to Ftarget does not belong to the DCS, the 2 PMs in the DCS surrounding Ftarget are applied in order to execute the task;
- if Ftarget is not available on the PE, the 2 PMs that surround Ftarget and which are in the DCS are applied.

A ring oscillator in STMicroelectronics 28nm FD-SOI technology has been simulated to validate the method proposed. In the examples (Fig. 1, Fig.2), 2 levels of body bias Vbb and 3 levels of supply voltage Vdd are available on the PE. The voltage values are: Vdd,1=0.6 Volt(V), Vdd,2=0.8V, Vdd,3=1.0V, Vbb_L=0V (unboosted state) and Vbb_H=1.2V (boosted).

Fig. 1 shows the frequency F in function of Vdd: F(Vdd).

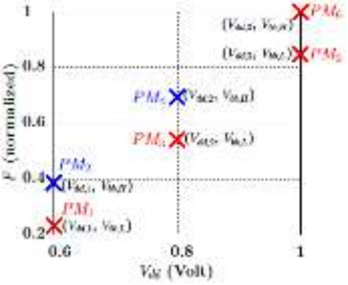


Figure 1 : Set of available frequencies F in function of the supply voltage Vdd.

Since the goal is to minimize power consumption, all the PMs are plotted in the P(F) plan (Fig. 2). Note that the PMs belonging to the DCS are plotted in red.

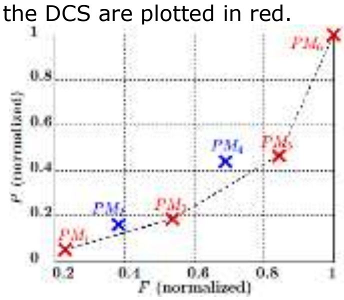


Figure 2 : Set of PMs in the (F, P) plane. Red crosses correspond to PMs in the discretely convex subset.

The power consumed when executing a task with Ftarget available on the PE but corresponding to a PM that does not belongs to the DCS is compared to the power consumed when executing the task with 2 PMs belonging to the DCS so as to obtain Ftarget on average. For instance, consider the target frequency Ftarget associated to PM2. Executing the task with PM1 and PM3 provides a gain in power consumption up to 27.55%.

Related Publications:
[1] Y. Akgul, D. Puschini, S. Lesecq, E. Beigne, P. Benoit, L. Torres, " Methodology for Power Mode selection in FD-SOI circuits with DVFS and Dynamic Body Biasing", IEEE Power and Timing Modeling, Optimization and Simulation (PATMOS), 2013.

FDSOI versus BULK CMOS at 28 nm node
Which Technology for Ultra-Low Power Design?

Research topics : Minimum energy operation, FDSOI, sub-threshold operation.

J. Mäkipää (VTT Finland), O. Billoint

ABSTRACT: To investigate benefits of utilizing FDSOI for ULP design, FDSOI and BULK CMOS 28nm nodes are compared by simulating a test circuit. Threshold voltage tuning by back-plane biasing (BPB) for FDSOI and bulk biasing (BB) for BULK is analyzed. Contours of constant energy with minimum energy points (MEPs). Simulation results show that FDSOI with forward BPB can be used effectively to control operating frequency and MEP operation. Implications of the results are discussed to give an overview how FDSOI performance gain over BULK CMOS can support in ULP design.

To explore the opportunity to utilize FDSOI technology, a BULK CMOS 28 nm process is compared to a state-of-the-art FDSOI CMOS 28 nm process from the same foundry using ring oscillator (RO) simulations. The results are shown for both strong and weak inversion operation, with a special emphasis on low power operation in sub-threshold and operating close to the minimum energy point (MEP).

The test circuit is a matrix of inverter gates connected as a ring oscillator and a number of delay chains, as shown figure 1. This known structure has already been studied in other papers to study influence of leakage current on performance and MEP position in a plane defined by supply voltage and substrate bias potential. We propose to apply this method to compare performances of two technologies of the same node, in order to study and quantify performance boost of FDSOI compared to BULK.

The first stage of the ring oscillator defines delay of the circuit, and accordingly maximum operation frequency. Stages from 1 to 9 emulate effect of the activity ratio α.

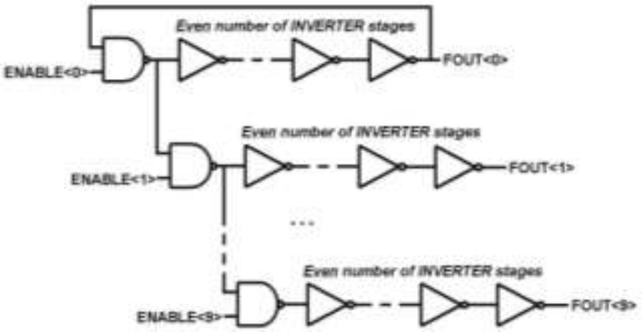

Even number of INVERTER stages
ENABLE<0>
FOUT<0>
Even number of INVERTER stages
ENABLE<1>
FOUT<1>
...
Even number of INVERTER stages
ENABLE<9>
FOUT<9>

Figure 1 : Test Circuit Topology

For an activity ratio of 0.1 all stages except stage 0 are deselected. Similarly, the activity factor can be varied from 0.1 to 1 with a step of 0.1 by activating a particular number of delay stages. The circuit was composed of standard nominal voltage gates of identical sizing for both technologies. Using the test circuit, data for contours of constant energy, relative operation frequency and energy delay product (EDP) were simulated in typical corner. Simulations were performed by sweeping supply voltage (Vdd), bulk bias (BB) and back-plane bias (BPB) voltage in order to vary effective threshold voltage.

Charge over one oscillation cycle was integrated and multiplied by supply voltage to get energy consumed by one cycle (or energy per operation, E/op). Period of oscillation was measured accordingly.

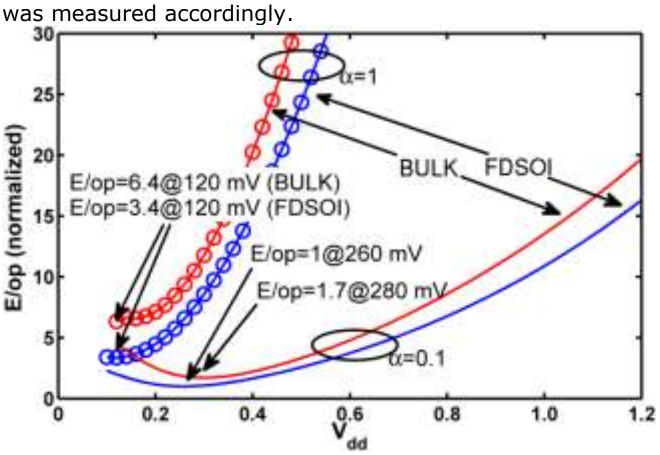


Figure 2 : Minimum energy points for BULK and FDSOI (no BB/BPB bias) with α=0.1 and 1.0

The E/op minimum energy point is located in supply versus BB/BPB voltage plane at a position where leakage and active energy are in balance. Leakage energy includes leakage current of all the ports and active energy consists of current consumed by active switching ports. Changing activity factor α modifies leakage to active energy ratio, thus varying α shifts the MEP position on the plane. Figure 2 shows position of the MEP (normalized to FDSOI energy per operation at Vdd=260mV) for both BULK and FDSOI in the typical corner without any BB or BPB modification. As observed, E/op of FDSOI is better than BULK across the whole Vdd range.

Simulation results indicate that FDSOI can help in pursuit of below-nominal supply voltage ULP operation. FDSOI consumes less energy at the MEP with higher operation frequency and allows for an extended power supply voltage range. Ultra-low power adaptive operation does not only enhance operation of current mobile devices but enables utilizing information processing in new application areas. Ambient intelligence, sensor networks, interactive environments and energy harvesting applications benefit from progress in ULP design.

Related Publications :

[1] Beigne, E.; Valentian, A.; Giraud, B.; Thomas, O.; Benoist, T.; Thonnart, Y.; Bernard, S.; Moritz, G.; Billoint, O.; Maneglia, Y.; Flatresse, P.; Noel, J.; Abouzeid, F.; Pelloux-Prayer, B.; Grover, A.; Clerc, S.; Roche, P.; Le Coz, J.; Engels, S. & Wilson, R. (2013), 'Ultra-wide voltage range designs in fully-depleted silicon-on-insulator FETs''Proceedings of the, 16th Design, Automation and Test in Europe Conference and Exhibition, DATE 2013, 18-22 March 2013, Grenoble', 613-618.

[2] Makipaa, J, Billoint, O., "FDSOI versus BULK CMOS at 28 nm node which technology for ultra-low power design?", 2013 IEEE International Symposium on Circuits and Systems (ISCAS), 19-23 May 2013, Bejing

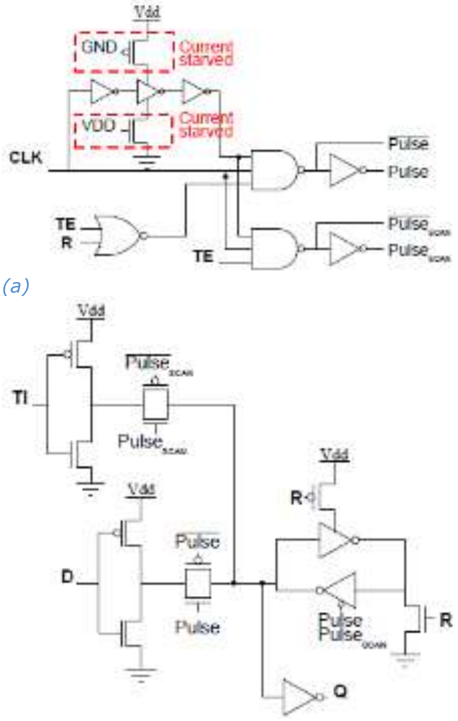
A Robust and Ultra-Low-Voltage Pulse-Triggered Flip-Flop in 28nm UTBB-FDSOI Technology

Research topics: CMOS digital circuits, flip-flop, Ultra-Wide Voltage Range

S. Bernard, A. Valentian, M. Belleville, D. Bol (UCL), J.-D. Legat (UCL)

ABSTRACT: So far, pulse-triggered flip-flops (pulsed-FFs) are mainly used in high-performance digital circuits, due to their small data-to-output delay. However, they suffer from a poor robustness to local variations occurring at ultra-low voltage. In this work, thanks to an innovative pulse generator, the operability of an energy-efficient pulsed-FF was extended in the ultra-low supply voltage range. Silicon measurements show that our pulsed-FF reaches a minimum operating supply voltage of 170mV, with a very small 6ns propagation delay at 250mV.

efficient and ultra-wide voltage range circuits for their small data-to-output delay. However, they suffer from a poor robustness to local variations occurring at ultra-low voltage. Pulsed-FFs are made of one latch open during a short period following the triggering clock edge (Fig. 1). This period is physically determined by a pulse signal, activating the latch and generated by a pulse generator. Due to local variations at ULV, the pulse signal may be shorter than the D-to-Q delay, and thus the flip-flop fails to latch the data.



(a)

(b)

Figure 1: Schematic of the Pulsed-FF: (a) the pulse generator, (b) the latch

The transistors of the latch were sized so as to reach a minimum energy-delay product. The novel pulse generator is based on a current-starved delay chain.

to variations, with a smaller hold time and a lower power dissipation [1].

In an industrial standard cell library, several functionalities are added to FFs at the transistor level. Additional transistors may impact the robustness of the FF at ULV. Therefore, our pulsed-FF also performs scan and reset functions in order to reach realistic complexity and robustness.

Silicon measurements were performed [2]. By setting the back bias voltages, GNDS and VDDS, at their nominal value, i.e. 0V, the pulsed-FF is found to be functional down to 250mV, with a propagation delay equal to 6ns. At 300mV and above, the propagation delay drops below 1ns, while state-of-the-art circuits at this supply voltage have an operating frequency in the MHz-range.

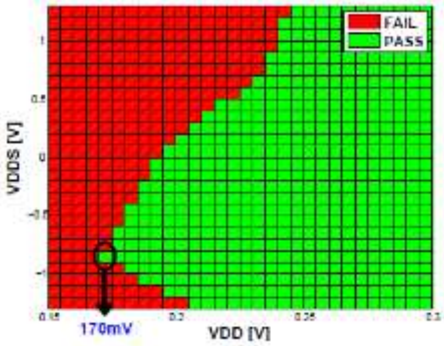


Figure 2: Pass/Fail graph showing the minimum operating voltage function of the supply voltage (VDD) and the PMOS back bias voltage (VDDS)

Although applying 0V on GNDS and VDDS leads to a balanced configuration at nominal supply voltage, the PMOS transistor becomes much weaker than the NMOS at ultra-low voltage. Therefore, a variation of the back bias of the PMOS transistors was performed to study the improvement of the minimum operating voltage. As can be seen on Fig. 2, a forward body bias (i.e. negative VDDS value) enhances the minimum operating voltage, while a reverse body bias (i.e. positive VDDS value) degrades it.

Related Publications:
1] S. Bernard, D. Bol, A. Valentian, M. Belleville and J.-D. Legat, “A Robust and Energy Efficient Pulse Generator for Ultra-Wide Voltage Range Operations,” 5th Asia Symposium on Quality Electronic Design (ASQED), Penang, Malaysia, August 2013
[2] S. Bernard, A. Valentian, M. Belleville, D. Bol and J.-D. Legat, “Design of a Robust and Ultra-Low-Voltage Pulse-Triggered Flip-Flop in 28nm UTBB-FDSOI Technology,” IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), Monterey, CA, USA, October 2013

Robust Clock Tree using Single-Well Cells for Multi-VT 28nm UTBB FD-SOI Digital Circuits

Research topics: Clock tree, UTBB FD-SOI, Multi-VT

B. Giraud, J.P. Noel (ST), F. Abouzeid (ST), S. Clerc (ST) and Y. Thonnart

ABSTRACT: The 28nm UTBB FD-SOI design platform enables multi-VT standard cells co-integration with independent body biases (BB). In this paper, we propose a new clock-tree cell to build a robust clock tree isolated from the various BB of the different VT regions. This cell uses mixed-VT transistors on p-Well, relying on an isolating shell of n-Well and deep n-Well to be integrated transparently in the implementation flows. This solution preserves propagation and transition times balancing improved by 2.5x, and a 5x drastic reduction of the clock skew at 0.4V compared to a conventional clock tree.

The 28nm UTBB FD-SOI (Ultra-Thin Body and BOX Fully-Depleted SOI) technology uses two VT flavors: regular VT (RVT) and low VT (LVT). For RVT logic gates, NMOS and PMOS lie on p-Well (PW) and n-Well (NW), respectively. The default biasing is GND and VDD for PW and NW, respectively. For LVT logic gates, NMOS and PMOS are based on NW and PW, respectively, both biased by default at GND.

Following the targeted performance at high and low voltages, different back biases (BB) are applied on the Wells to increase speed or reduce leakage with different co-integration strategies. For Back Biasing capability on both NW and PW, and in both RVT and LVT, standard cells must be grouped by blocks within a deep NW (DNW), as depicted in [1], leading to a large area overhead.

A key feature of digital circuits is the implementation of the clock tree. For delay and slope balancing, a single VT flavor is usually used for the clock tree implementation, and skew is fine-tuned by MOS sizing. In UTBB FD-SOI, when both RVT and LVT are used, a single VT clock tree would lead to huge area overhead because of DNW gaps, while mixed VT clock tree would degrade the clock skew. This skew would be worsened with back biasing on surrounding standard cells whose delay would be modified differently according to their VT.

To overcome this issue, we have proposed a clock tree insensitive to BB modulation of surrounding logic cells, using clock-tree cells isolated in a PW island surrounded by NW & DNW inside a double-height single PW (SPW) cell (Fig. 1). Thus, the clock path becomes an independent BB region.

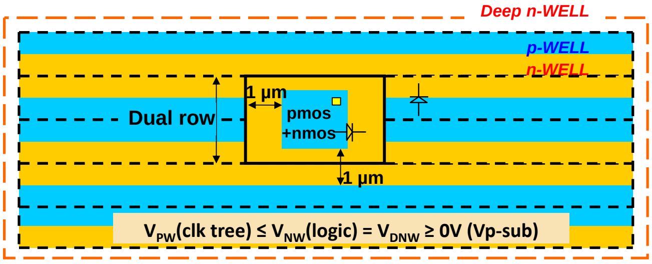

Deep n-WELL
p-WELL
n-WELL
Dual row
1 µm
pmos
+nmos
1 µm
V_{PW}(clk tree) ≤ V_{NW}(logic) = V_{DNW} ≥ 0V (Vp-sub)

Figure 1: Proposed double-height SPW clock-tree cell insensitive to BB of surrounding data gates

SPW clock-tree cells are intrinsically unbalanced because they consist of LVT PMOS and RVT NMOS. To rebalance the rise/fall times, PMOS VT can be increased using the short channel effect. Therefore, gate length of PMOS has been increased by 4nm.

Table I shows an improvement of 2.5x on difference between rising & falling edges propagation time and transition time compared to the existing cell when BB on surrounding data cells varies from 0V to 2V. The proposed solution preserves the clock tree balance.

To illustrate the skew improvement, an in-house 200k Gate 32b DSP core was implemented in 28nm UTBB FD-SOI technology up to the place & route, resulting in distinct areas dedicated to RVT and LVT. As a next step, two options of 13k flip-flop clock tree synthesizes were considered: one mixing RVT/LVT clock tree cells sharing their Well with surrounding logic cells and one using the proposed SPW solution with fully isolated Wells.

Performance targets for these clock trees have been set to 1.6GHz in typical conditions, with a clock skew of less than 10% of the clock period. Post-layout extraction has been performed to run Spice simulations of the clock propagation delays to all the leaves in various VDD/BB conditions.

Table I shows the resulting skew increase while reducing voltage and/or high forward BB (not represented here) on a mixed RVT/LVT clock tree. On the contrary, SPW clock tree, with an independent BB voltage always maintains clock period and clock skew in the same ratio whatever the supply & BB conditions, around 8% for this DSP. Compared to the existing solution, this results in a gain of 50%, 3x and 5x at 1V, 0.7V and 0.4V, respectively.

Table I: Result summary

	Δ propag. time (norm.)	Δ rise/fall time (ps) (norm.)	Max. deviation of delay vs. VDD:		
			0.4V	0.7V	1V
Existing solution	1	1	46%	20.1%	11.6%
Proposed solution	0.2	0.2	8.7%	6.9%	7.4%

Related Publications:
[1] B. Giraud, J.P. Noel, F. Abouzeid, S. Clerc and Y. Thonnart, " Robust Clock Tree using Single-Well Cells for Multi-VT 28nm UTBB FD-SOI Digital Circuits", IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2013.

Optimization of a Voltage Sense Amplifier operating in Ultra Wide Voltage Range with Back Biasing in 28nm UTBB FD-SOI

Research topics: SRAM, Sense Amplifiers, UTBB FD-SOI

G. Moritz, J. -P. Noel (ST), B. Giraud, A. Grover (ST), D. Turgis (ST)

ABSTRACT: Advanced SoC designs regularly use Dynamic Voltage and Frequency Scaling (DVFS) to achieve both high frequencies and low dynamic power targets in portable systems. In this study, we focus on optimization of a Voltage Sense Amplifier (VSA) in 28nm Ultra-Thin Body and BOX Fully Depleted SOI (UTBB FD-SOI) technology to achieve high performance operations over the Ultra Wide Voltage Range (UWVR) from 1.3V to 0.4V. We use forward body bias modulation to extend operation range of the VSA and also reduce sense amplifier read time by 28%, while saving power consumption by up to 59% compared to Bulk.

While achieving high frequencies is necessary for application processors, high energy efficiency is also clearly a key differentiator in the context of portable electronics with limited battery life time. To deal with this tradeoff, design techniques like Dynamic Voltage and Frequency Scaling (DVFS) are commonly used. High speed access of SRAMs across an Ultra Wide Voltage Range (UWVR) is essential to ensure processor efficiency in a DVFS implementation. Correct and fast read operation mainly relies on a voltage sense amplifier (VSA) circuit and its input voltage difference transferred by the bit-lines (offset).

Figure 1 shows that the conventional VSA designed in UTBB FD-SOI technology enables correct read operations with small offsets at low supply voltages.

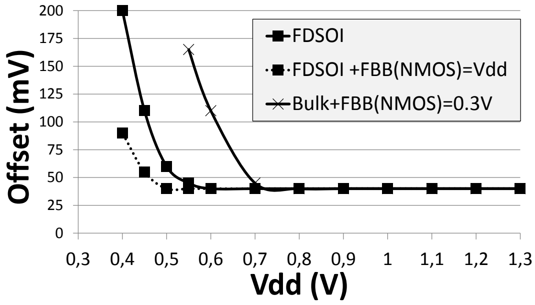


Figure 1: Worst VSA offset variation versus Vdd for UTBB FD-SOI with FBB=0V, FBB=Vdd and for Bulk technology with FBB=0.3V

Actually, in Bulk technology, it reaches limits around 0.6V supply voltage with 110mV offset compared to only 40mV in UTBB FD-SOI technology. This reduction of 64% directly impacts the power consumption and the maximum SRAM operating frequency.

Therefore, better improvements appear by using the UTBB FD-SOI degree of freedom. The 'back gate' enables wide voltage range for body biasing. Forward body biasing (FBB) can be easily applied to NMOS (on NWELL) VSA's without adding a deep NWELL. By increasing FBB up to 1V, VSA can operate at 0.4V supply voltage with an offset of 85mV.

Low process variations of UTBB FD-SOI is another advantage which confirm such enhancement and that the technology is suitable for very low supply voltage applications such as SRAM.

Table I shows that power consumption is 50% lower without FBB and up to 60% lower with FBB compared to bulk for the same read time.

Table I a) Total power at given VSA read time and b) VSA read time at given total power

a)	ref. Bulk (FBB=0.3V)	UTBB FDSOI (FBB=0V)	UTBB FDSOI (FBB=Vdd)
time	**power (Vdd)**	**power (Vdd)**	**power (Vdd)**
56ps	113µW (0.9V)	57µW (0.8V) → -50%	47µW (0.7V)→ -59%
b)			
power	**time (Vdd)**	**time (Vdd)**	**time (Vdd)**
110µW	55ps (0.9V)	45ps (0.9V) → -19%	40ps (0.9V)→ -28%

Figure 2 exposes FBB optimization points. For very low offset, an FBB increase results in a higher error rate. It appears that an optimal couple of offset and FBB exists to reach no reading error at very low supply voltage.

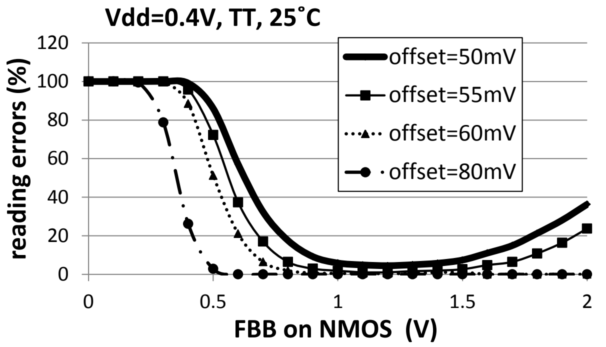


Figure 2: Reading error percentage versus FBB applied on NMOS of VSA design in UTBB FD-SOI technology

We have demonstrated that applying extended FBB can further enhance the benefits of UTBB FD-SOI technology in efficiently increasing performance of one of the most critical circuits in a UWVR SRAM. A large FBB range can be used to greatly improve the minimal operating supply voltage down to 0.4V. By coupling UTBB FD-SOI technology with the proposed design methodology, VSA read time can be reduced by 28% while simultaneously reducing power consumption by up to 59% compared to Bulk technology. The proposed optimization techniques can also be applied to most sense amplifier designs.

Related Publications:
[1] G. Moritz, J. -P. Noel, B. Giraud, A. Grover, D. Turgis, «Optimization of a Voltage Sense Amplifier operating in Ultra Wide Voltage Range with Back Bias DesignTechniques in 28nm UTBB FD-SOI Technology», ICICDT, 2013.

SRAM row decoder design for Wide Voltage Range in 28nm UTBB-FDSOI technology

Research topics: memory, UTBB FD-SOI, Single Well

Grégory Suraci, Bastien Giraud, Thomas Benoist, Adam Makosiej, Olivier Thomas

ABSTRACT: This paper focuses on the design of SRAM row decoder for modern portable devices, in 28nm Ultra-Thin Body and Buried oxide (UTBB) Fully-Depleted SOI (FDSOI) technology. The proposed Mixed Single Well (Mixed-SW) design concept enables a major speed improvement over a wide voltage range with no standby power penalty, as compared to a regular VT (RVT) design. The simulation results of a Mixed-SW dual-port SRAM row decoder show 16% and 43% propagation delay reduction at 1V and 0.6V, respectively. The gain obtained at RVT design standby power is enabled by the wide range N-Well back biasing.

High-speed and low-power requirements in modern portable devices demand the circuit to operate over a wide range of supply voltages (VDD) to maximize the operation time on battery. The UTBB-FDSOI technology enables the reduction of the minimum operating VDD of SRAM array. However, tracking logic speed at low voltage remains a major challenge for SRAM. This work focuses on the design of a low-voltage, high-speed and low-standby-power SRAM Row Decoder. The proposed circuit design approach, called Mixed Single-Well (Mixed-SW), is detailed. The obtained performance gains are demonstrated on a 28nm SRAM row decoder for dual-port bitcell based memory.

UTBB-FDSOI devices are implemented on an ultra-thin silicon layer on top of a buried oxide (BOX). The thin BOX isolation enables the application of wide range back biasing with no source/drain-substrate junction leakage and the use of either N-well (NW) or P-well (PW) for both NMOS and PMOS devices. PW NMOS devices associated with NW PMOS devices provide regular VT (RVT) logic gates. By flipping and grounding the wells, low VT (LVT) logic gates can be obtained. In addition, a Single-NW (SNW) or Single-PW (SPW) design offers hybrid {RVT, LVT} logic gates, providing an intermediate time propagation delay and standby power consumption.

The proposed Mixed-SW design concept consists in interleaving SPW and SNW-based logic gates (Fig. 1). A SNW or SPW logic gate is placed in a logic path depending on the output logic state in standby mode to minimize the static power consumption. For instance, when a low logic level is expected as a dominant output a SNW gate can be used resulting in leakage through a RVT P-network.

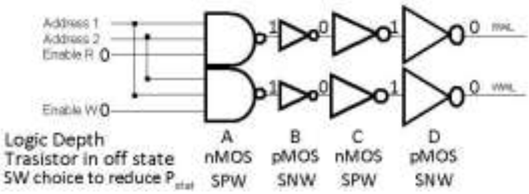


Figure 1: Choice of the Well type on dual-port SRAM Mixed-SW row decoder in standby mode

In the operating mode, the NW can be grounded (VNW = VPW = gnd) to lower the VT of the RVT PMOS devices (SRVT ≈ RVT–60mV @1V) in single NW, leading to reductions of both low-to-high and high-to-low propagation delays.

Compared to RVT, the average Mixed-SW propagation delay approaches LVT performance gains. With VDD scaling, the gain increases even more, reaching 43% at 0.6V versus 56% in LVT. In standby mode, the NW is biased at VDD, making the row decoder leak as low as the RVT, while the LVT consumption is at least 16x higher.

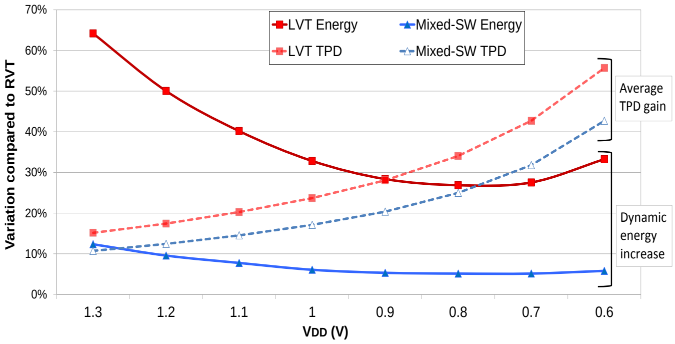


Figure 2: Row decoder gains of time propagation delay and energy consumption compared to equivalent RVT design (TT@27°C, VPW = gnd, VNW = gnd)

The Mixed-SW design concept provides an alternative for high-speed, low-voltage and low-standby-power SRAM row decoder, particularly attractive for embedded SRAM in mobile device processors. Optimum energy efficiency is achieved both in operating and standby modes owing to the back bias control.
The described technique was implemented in a full SRAM macro and is expected to be validated in silicon measurements.

This approach can be extended to any buffer or logic path in SRAM and other custom designs, in which the designer can identify the dominant logic states in standby.

Related Publications:
[1] G. Suraci, B. Giraud, T. Benoist, A. Makosiej and O. Thomas, " SRAM row decoder design for Wide Voltage Range in 28nm UTBB-FDSOI", IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2013.

Digital Circuits in 28nm UTBB FD-SOI: Greater Gains with larger Body Biasing Range

Research topics: UTBB FD-SOI, energy efficiency, ultra-wide voltage range

B. Giraud, J.-P. Noel (ST), P. Flatresse (ST), O. Thomas

ABSTRACT: The first digital circuit in 28nm UTBB FD-SOI has been designed including 3 LDPC decoders that are benchmarked against 28nm Bulk over an ultra wide voltage range. Thanks to FD-SOI isolation, we demonstrate that the well-known Bulk back bias range of -/+300mV can be significantly extended by 5X enabling to outperform FD-SOI intrinsic gains up to 49% in power saving or 35% in speed increase with respect to 28nm LP Bulk technology at 1V. The proposed FD-SOI architecture with dual STI can achieve 10X wider back bias range while supporting both forward and reverse back bias, in order to fully exploit the possibilities of this technology.

UTBB (Ultra-Thin Body and BOX (Buried OXide)) FD-SOI technology exhibits excellent electrostatic control and low variability. Compared to Bulk, this technology demonstrates better performances (greatly improved with Vdd reduction) with similar leakage current at same Vdd.

The great advantage of FD-SOI stems from the coupling capacitance with the back interface. Therefore, the threshold voltage can be adjusted roughly by doping n- or p-type the BP (Back Plane: over-doped area located below the BOX) and finely by biasing the BP.

Fig. 1 represents Well configurations in Bulk and in FD-SOI. The BB (Back-Bias) permitted range which is limited to +/-300mV in Bulk, is extended by 5X in FD-SOI with single STI (Shallow Trench Isolation) thanks to the BOX isolation. About -3V or +3V is possible for conventional Well (CW) and flip Well (FW), respectively (Fig. 1).

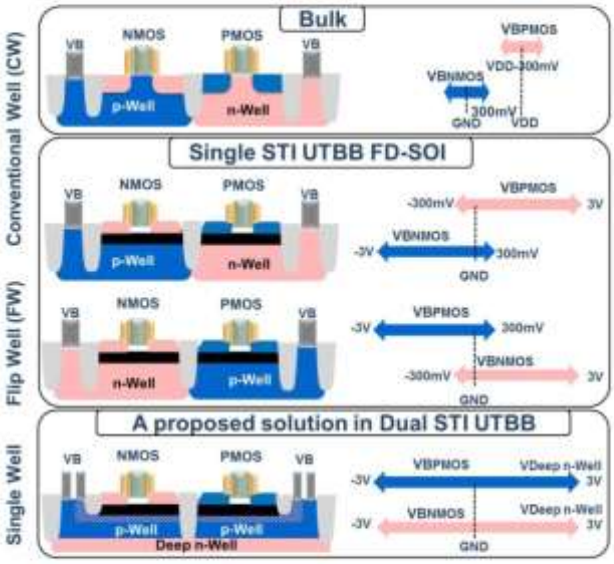


Figure 1: Well layer configurations and BB permitted range of Bulk, FD-SOI with single and dual STI [1,2]

[1] presents the comparison of 3 LDPC (Low-Density Parity-Check) decoders in Bulk and FD-SOI for the same technology node (28nm) and the same design (same RTL, synthesis and Place and Route).

Fig. 2 points out the better energy efficiency of LDPC in FD-SOI. The same frequency can be achieved by reducing Vdd by 200mV with no BB and by 300mV with 1V Forward BB (FBB), leading to 49% reduction of total power compared to 1V Bulk. Similarly, the same power can be obtained by reducing Vdd in FD-SOI by almost 200mV. However, the FD-SOI frequency is still higher (35%) than Bulk.

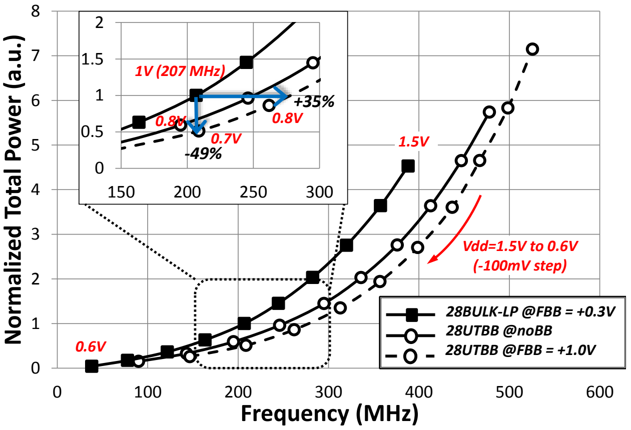


Figure 2: Measurement of total power vs. frequency [1]

Besides, in Bulk, the minimum operating Vdd of 600mV can be greatly reduced down to 410mV and 350mV in FD-SOI with no BB and with 1V FBB, respectively.

BB, the key feature of UTBB FD-SOI enables better performances and lower total power. [2] proposes a promising solution using an optimized Well isolation thanks to the dual STI and the dual Well/BP tap (Fig. 1). Well junction removed by dual STI enables both reverse BB and FBB (impossible in single STI). Furthermore, the BB range can be further increased (2X compared to single STI), leading to optimal energy efficiency over a maximized Vdd range.

Related Publications:
[1] P. Flatresse, B. Giraud, J.-P. Noel, et al., "Ultra Wide Body Bias Range LDPC decoder in 28nm UTBB FDSOI Technology", ISSCC, San Francisco, February 2013
[2] L. Fenouillet, M. Vinet, J. Gimbert, B. Giraud, J.P. Noël, et al., "UTBB FDSOI transistors with dual STI for a multi-Vt strategy at 20nm node and below", IEDM, San Francisco, December 2012.

4

Circuits in Emerging Technologies

NanoRelays
Resistive Switching Devices
Ambipolar Devices

Interest of using NEMS switches for implementing adiabatic logic circuits

Research topics: Adiabatic Circuit, Nano-electromechanical switch

S. Houri, A. Valentian, H. Fanet and C. Poulain

ABSTRACT: Along with the need to embed electronic systems into the surrounding environment, such as embedded sensors, processors, and communication nodes, comes the need for an ever lower power consumption per operation in logic circuits.

We have studied the impact of using nano-electromechanical switches on the power dissipation of adiabatic logic circuits. Compared to CMOS-based ones, it is shown that NEMS-based adiabatic circuits offer the distinctive advantage of zero static losses, leading to expected energy dissipation orders of magnitude lower.

Electrostatic nano-electromechanical (NEM) switches have already been suggested and demonstrated for use in classical logic circuits. NEM switches offer the advantage of zero leakage current and therefore zero static power dissipation, which is an appealing property for low power low performance circuits. However, these switches require high operating voltages and suffer from low switching speeds when compared to MOSFETs: these factors remain the major drawbacks that nullify any advantages they may offer. Nonetheless, NEM switches are ideal candidates to replace classical CMOS elements in adiabatic logic circuit applications, whereby using adiabatic charging, it is possible to offset the high energy dissipation that accompanies the voltages required to operate them [1].

Logic operations at circuit level may be reduced to either charging or discharging a capacitor through a controlled switch, i.e. a transistor. When charging is done by a constant voltage source, as is the case with classical circuits, a charge-discharge cycle would always dissipate an energy per cycle equal to $C*Vdd^2$, irrespective of the resistance value of the switch (where C is the output capacitance and Vdd the supply voltage).

However, when adiabatically charging the capacitor, i.e. charging with a voltage ramp, it can be shown that the energy dissipated in a charge-discharge cycle is equal to:

$$E = 2\frac{RC}{T}CV_{dd}^{2}$$

where R is the switch resistance and T is the ramp duration, as shown schematically in Fig. 1. The above expression is a valid approximation for T/RC>10.

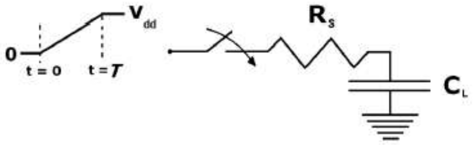


Figure 1: Schematic representation of adiabatic charging of an RC circuit

NEM switches of varying design, dimensions and materials have been constructed and demonstrated. In this work a 1-dimensional generic model was developed. It can be applied to all devices equally in order to obtain a generic formulation of the energy efficiency and performance of NEMS-based adiabatic logic circuits [2][3].

Using this model, the total energy dissipation of an adiabatic logic circuit can be simulated: it is represented by the solid red line in Fig. 2 [4]. It can be seen that there is no minimum energy point, as NEM switches are leakage free. The analysis of CMOS-based adiabatic logic circuits was also done for comparison purposes: in that case, there is an optimal energy point, obtained when the leakage power is equal to the switching power.

A figure of merit (FOM) Energy*delay was defined: it is plotted in dashed lines in Fig. 2. It shows that for both devices, once a certain value is attained, the FOM tends to an asymptotic value.

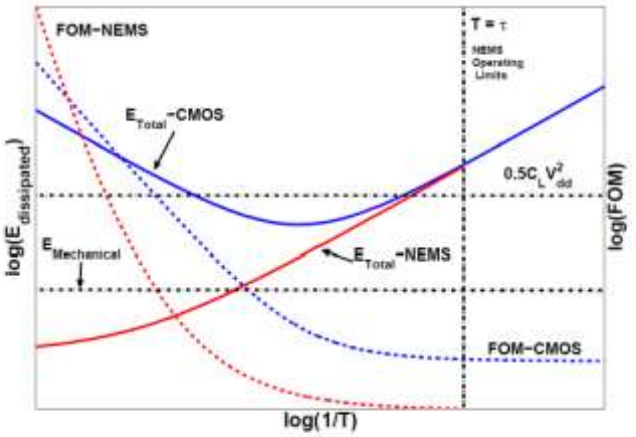


Figure 2: Plots showing the total energy dissipation (solid lines) and the FOM (dashed lines) for CMOS (blue color) and NEMS based (red color) adiabatic logic circuits

Related Publications:
[1] H. Fanet, "Circuit logique à faible consommation et circuit intégré comportant au moins un tel circuit logique", Patent n° 11 56670
[2] S. Houri, A. Peschot, C. Poulain, A. Valentian and H. Fanet, "In-situ Characterization of Gold contacts in RF MEMS Switches," 10th International Workshop on Nanomechanical Sensing (NMC 2013), Stanford, California, USA, May 2013
[3] S. Houri, A. Valentian, H. Fanet and C. Poulain, "Performance envelope of Adiabatic Logic Circuits Based on Electrostatic NEM Switches," 12th Edition of IEEE Faible Tension Faible Consommation, Paris, France, June 2013
[4] S. Houri, A. Valentian and H. Fanet, "Comparing CMOS-Based and NEMS-Based Adiabatic Logic Circuits," 5th Conference on Reversible Computation, Victoria, Canada, July 2013

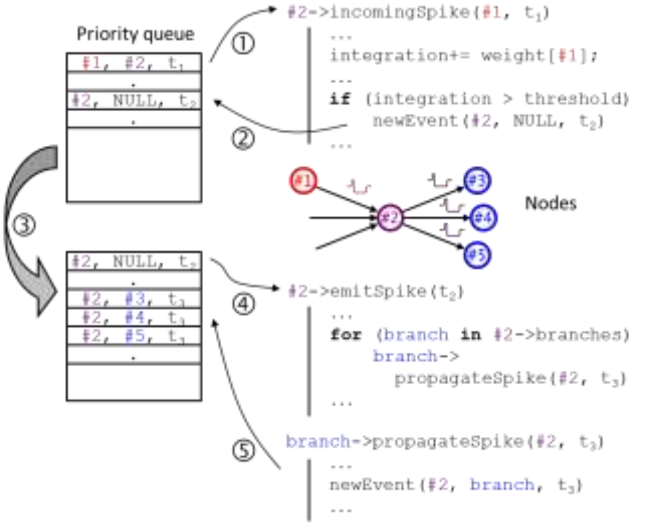
Design Exploration Methodology for Memristor-Based Spiking Neuromorphic Architectures with the Xnet Event-Driven Simulator

Research topics : Xnet, event-driven simulator, memristor, spike-based, neural network

O. Bichler, D. Querlioz (IEF, CNRS), D. Roclin, C. Gamrat

We introduce an event-based methodology, and its accompanying simulator ("Xnet" [1]) for memristive nanodevice-based neuromorphic hardware, which aims to provide an intermediate modeling level, between low-level hardware description languages and high-level neural networks simulators used primarily in neurosciences. This simulator was used to establish several results on Spike-Timing-Dependent Plasticity (STDP) modeling and implementation with Resistive RAM (RRAM), Conductive Bridge RAM (CBRAM) and Phase-Change Memory (PCM) type of memristive nanodevices.

Following the advancements in computational neuroscience, spiking neuromorphic hardware has gained momentum over the last years. This trend is reinforced with the latest proposals to use memristive nanodevices as synapses, which are particularly attractive to implement efficient timing-based learning rules like Spike-Timing-Dependent Plasticity (STDP) in dense crossbar arrays. A major focus is to capture biological processes with a much higher realism than earlier Artificial Neural Networks (ANN), thus enabling richer interactions with neuroscience, large-scale hardware accelerated neural simulations and real-time behaving systems. Another emerging field of applications for SNN are hardware Intellectual Property (IP) cores, especially in embedded computers. SNN could indeed complement or replace otherwise computationally heavy sensor processing, like audio or video patterns extraction, learning, recognition and tracking.



```
#2->incomingSpike(#1, t1)
    ...
    integration+= weight[#1];
    ...
    if (integration > threshold)
        newEvent(#2, NULL, t2)
    ...
```

```
#2->emitSpike(t2)
    ...
    for (branch in #2->branches)
        branch->
            propagateSpike(#2, t3)
    ...
```

```
branch->propagateSpike(#2, t3)
    ...
    newEvent(#2, branch, t3)
    ...
```

Figure 1: Events processing in Xnet. The next scheduled event is processed by its destination in incomingSpike() (1), which can create internal events (2) that are inserted into the priority queue (3). Internal events are processed by emitSpike() (4), which emit events towards outputs by calling their propagateSpike() method (5).

To model such systems, hardware description languages do not provide the appropriate level of abstraction for fast and efficient architectural exploration, while neural network simulators popular in the neuroscience community lack the integration of synaptic memristive device modeling. To provide an intermediate modeling level for neuromorphic hardware, we introduce several event-based simulation strategies that are implemented in our event-driven simulator ("Xnet"). The Xnet event processing engine is presented in figure 1. It was developed to allow fast and efficient design exploration of spiking neuromorphic architectures by providing a framework that can mix high-level behavioral modeling with hardware constraints integration. More specifically, it was designed with spiking retina (one of the many possible input stimuli, see figure 2) and memristive nano-devices based architectures in mind.

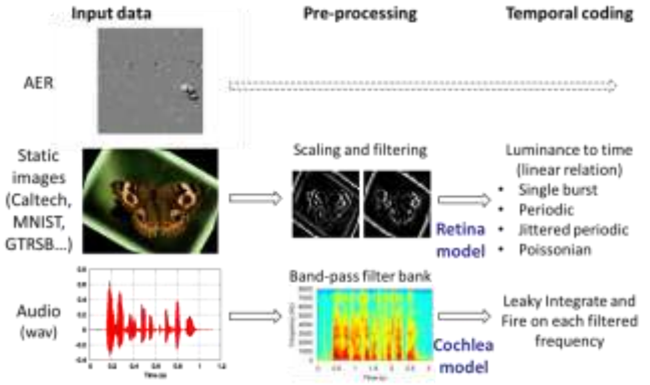


Figure 2: Possible input stimuli: AER recording, static image and audio waveform. The pre-processing and temporal coding steps can be adapted to emulate hardware sensors or pre-processing units.

This simulator, along with its accompanying methodology was used for the simulations in several papers published by our group [2], [3]. We hope that Xnet will be a first step towards a generic and efficient framework for high-level memristor-based spiking neuromorphic architectures exploration. It proved to be extremely valuable for rapid evaluation of new nanodevices and for a better understanding of STDP-like learning rules. The Xnet source code is currently not licensed, but is available through partnership with CEA LIST. Additional information on implementation details is available upon request.

Related Publications:
[1] O. Bichler, D. Querlioz, D. Roclin, C. Gamrat, "Design Exploration Methodology for Memristor-Based Spiking Neuromorphic Architectures with the Xnet Event-Driven Simulator", Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on
[2] D. Roclin, O. Bichler, C. Gamrat, S.J. Thorpe, J.-O. Klein, "Design Study of Efficient Digital Order-Based STDP Neuron Implementations for Extracting Temporal Features", Neural Networks (IJCNN), The 2013 International Joint Conference on
[3] M. Suri et al., "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction", Electron Devices Meeting (IEDM), 2011 IEEE International

Spiking Neural Network for Embedded image sensors

Research topics: Embedded Spiking Neural Network, Nano-Devices, Memristors, CBRAM

D.Roclin, O.Bichler, C.Gamrat, J-O.Klein (IEF paris-XI)

ABSTRACT: Our work focuses on the implementation of a Spiking Neural Network (SNN) for an embedded image sensor. To make it embeddable, there are 2 major constraints: power and area. We first analyzed each of the three parts of a SNN, the synapse, the neuron and the learning method, in order to optimize its implementation in a digital architecture and to gain in silicon area and power consumption. To further improve this gain, we chose to implement the synapses of the SNN with emergent nano-devices. Usually realized with flash memory, we wanted to study the gain by designing the first neuromorphic chip with CBRAM devices as synapses.

In a previous work [3], we have shown that Spiking Neural Network (SNN) is an effective mean for processing large natural data sets. The aim of this work is to investigate possibilities for embedding such a network in an image sensor in order to retrieve high level information at the sensor output. We analyzed the main parts of the SNN architecture: synaptic array, neuron and learning mechanism, with respect to silicon area and power consumption. We used an event-based simulator for the task and we started from a previously established simulation, which emulates an analog spiking neural network that can extract complex and overlapping, temporally correlated features in a vision application.

Starting at the synapse level, our first objective was to quantize the lowest bit resolution that maintains the learning ability to and the detection of visual features. We first demonstrated that the SNN was still functional with only 2-bit synaptic weights. Such a bit resolution saves area and energy in a digital architecture, or reduces the need for multi-level storage in a memristor based analog architecture.
Neurons in the SNN circuits are generally of the Leaky Integrate and Fire (LIF) type; hence implementing the leak is a major component of the neuron behavior. Indeed an exponential leak on a digital system is space consuming and adds latency due to the calculation complexity. Therefore, we proposed to implement a linear leak to eliminate these costs. This change eliminated the exponential calculation and replaced it with a multiplicative calculation. This lowered the area and latency of the Arithmetic Logic Unit of the system. We proved that a linear leak did not affect the learning capability [1].

Spike-Timing-Dependent-Plasticity (STDP) or a derivative is generally used as the learning paradigm in SNN. The STDP mechanism requires the implementation of a timestamp for each input neuron that stores the time of the last spike. When an output neuron fires, it compares each synaptic activation timestamp to the LTP window time and decides if the synapse will undergo a Long-Term-Potentiation (LTP) or a Long-term-Depression (LTD). This implementation is space consuming due to the large number of required comparators and counters. Our hypothesis is that the order of arrival of the presynaptic spikes is as important as their precise arrival time. We proposed to replace the time-based LTP window with a First In First Out (FIFO) buffer. When an output neuron fires, it performs a LTP on the synapses present in the FIFO and a LTD on all the others.
This work showed that a FIFO of 200 events is already efficient. This implies that implementing a FIFO memory shared by all output neurons is a valid solution [1].

To further improve the area and power consumption of the SNN we chose to replace the common used Flash memory that implement the synapses with emergent Conductive-Bridge RAM (CBRAM) memristors. This technology has many benefits as for example its non-volatility (no need to refresh the stored data and it is kept even if the power supply is turned off), its low programming power consumption, its scalability and its compact structure when arrange in matrix (1T-1R) or crossbar (1R).

We designed a test chip that includes a 4*4 CBRAM matrix and a 4*4 CBRAM crossbar as well as single CBRAM devices (Fig.2). The chip also includes CMOS circuits that drive the different structures making the chip controllable with digital inputs.

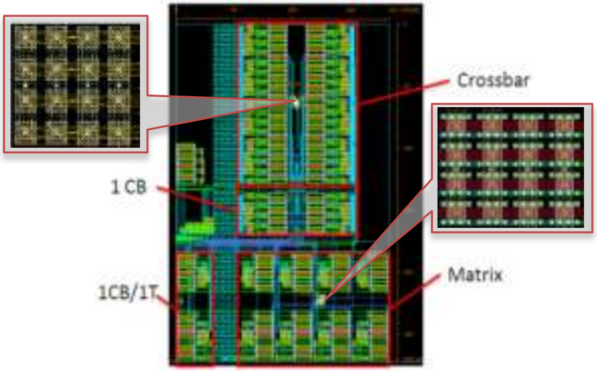


Figure 2: Test chip Layout with crossbar and matrix of CBRAMs

Our neuromorphic hybrid CMOS/CBRAM test chip will be used for experimental characterizations of SNN circuits:
- to validate the use of memristive devices such as CBRAM for implementing synapses,
- to retrieve power consumption data,
- to test various learning schemes for SNN including stochastic techniques,
- to evaluate the difference in area and power consumption compared to synapses implemented with standard CMOS.

Related Publications:
[1] D.Roclin et al. “Design Study of Efficient Digital Order-Based STDP Neuron Implementations for Extracting Temporal Features”, International Joint Conference on Neural Networks (IJCNN), Dallas, USA, 4-9 August 2013
[2] D.Roclin et al., 'CBRAM as synapses for neuromorphic engineering'. The 2013 NanoSaclay nanoelectronics international workshop
[3] O. Bichler et al. “Extraction of temporally correlated features from dynamic vision sensors with spike-timing-dependent plasticity,” Neural Networks, vol. 32, pp. 339 – 348, 2012.

Self-Checking Ripple-Carry Adder with Ambipolar Silicon NanoWire FET

Research topics : self-checking, ambipolar, SINW, online-testing

O. Turkyilmaz, F. Clermidy, L. G. Amarù(EPFL), P.-E. Gaillardon(EPFL), G. De Micheli(EPFL)

ABSTRACT: For the rapid adoption of new and aggressive technologies such as ambipolar Silicon NanoWire (SiNW), addressing fault-tolerance is necessary. Traditionally, transient fault detection implies large hardware overhead or reduced performance compared to permanent fault detection. In this paper, we focus on on-line testing and its application to ambipolar SiNW. We demonstrate that ambipolarity on a self-checking ripple-carry adder can help reduce the hardware overhead. In comparison to equivalent CMOS process, ambipolar SiNW design shows 56% (28%) less area with 62% (6%) reduced delay for Static (Transmission Gate) design style.

Ultimate CMOS technology leads towards 1D devices. These devices have good channel control properties and limited fabrication complexity. Among 1D devices, Silicon Nanowires (SiNWs) have strong arguments thanks to classical CMOS material compatibility. Similar to some of the 1D structures, they present ambipolar behavior, i.e., both n- and p-type conduction, that can be controlled by the use of an extra gate (Fig.1). Recently, very efficient implementations of digital circuits, e.g., XOR, have been demonstrated using this technology and more specifically its ambipolar property.

Extreme scaling and increased operation frequencies, lead devices to deeper submicron levels with lower noise margins, requiring a solution for fault tolerance. In this context, online testing offers one of the most adequate solutions for robust circuits.

Online testing enables the detection of temporary and permanent faults immediately after the fault occurs. In comparison to offline testing, it does not require stalling the system operation for error diagnostics and thus eliminates complex software routines to test the unit. Redundancy is the most general approach which replicates the blocks temporally or physically to detect faults. Alternatively, self checking circuits provide an efficient solution for testability which checks the results continuously for transient and permanent faults and thus prevents data contamination.

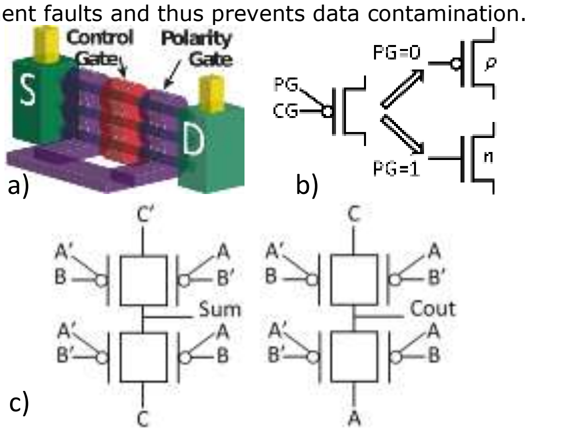


Figure 1: Ambipolar SiNWFET a)structure b)polarity control c)full adder

In this work [1], a Self-Checking Ripple Carry adder in Ambipolar SiNWFET technology with Carry-Checking/Parity-Prediction scheme is presented. We take advantage of very efficient implementation of XORs in ambipolar SiNWFET to reduce the implementation cost due to the high number of XOR-related operations used in self-checking adders.

A full adder is designed firstly to observe the advantage of ambipolar SINWFET technology (Fig.1c). Results show that it is 25% faster than TG-CMOS and 64% than the static CMOS while requiring 71% and 43% less area than its static and TG-CMOS counterparts, respectively.

An example of 4-bit self-checking adder with parity prediction and checking is shown in Fig. 2a. Complemented carries are checked against the real ones using the double rail (2R) checkers (Fig. 2b). The faults on the carry signals are detected using the double rail checkers because all errors are propagated through the checker tree. If no error is propagated through the carries, only one sum output is affected from the fault. This error is detected by the parity prediction circuitry. The final adder uses 1-bit adder blocks (Fig. 2c) which generate the carry, sum and the complemented carry. This 1-bit adder can be designed with one full adder and complemented carry generation circuitry.

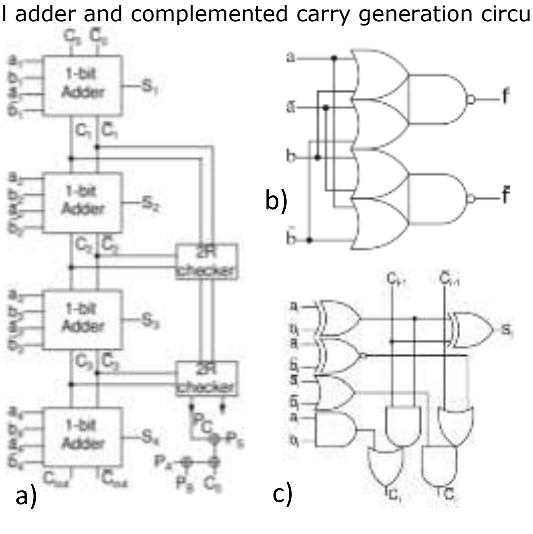


Figure 2: a) Self-Checking 4-bit Ripple Carry adder using Carry-Checking/ Parity-Prediction scheme b) Double rail checker c) 1-bit adder with complementary carry generation

We evaluated the area and delay of the proposed in Static, TG-CMOS and Ambipolar SiNWFET technologies with bit widths ranging from 4 to 128. The results show that the new adder is at least 56% (28%) smaller and 62% (6%) faster than its Static CMOS (TG-CMOS) counterparts.

Related Publications:

[1] Turkyilmaz, O.; Clermidy, F.; Amaru, L.; Gaillardon, P.-E. & De Micheli, G. (2013), "Self-checking ripple-carry adder with Ambipolar Silicon NanoWire FET", Proceedings of the, 2013 IEEE International Symposium on Circuits and Systems, ISCAS 2013, 19 May 2013 through 23 May 2013, Beijing', 2127-2130

5

PhD Degrees Awarded in 2013

OGUZ Alp

BELMAS François

CHRISTMANN Jean-Frédéric

JOUBERT Antoine

GUTHMULLER Eric

VINCENT Lionel

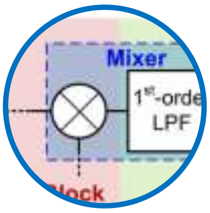
BERTOLINI Clément

CARBON Alexandre

STAN Oana

ELSAHMARANY Lola

PhD degrees awarded in 2013



OGUZ Alp
University: EPFL

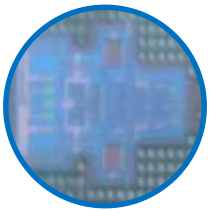
Development of Reconfigurable Circuits for WPAN Applications

This thesis presents the development of a low-power analog-to-digital converter (ADC) which is intended to be used in WPAN receivers. One common problem in WPAN systems is that the receiver, operates as over-specified when the conditions are not demanding. Other problems of WPANs include market segmentation and high design margins due to process variations. All these problems result in reduced power efficiency.

In this work, an adaptive reconfiguration concept is proposed. Adaptive reconfigurability is based on the idea that environment-aware systems can prevent the power inefficiencies by reconfiguring themselves to the desired level of performance. Moreover, in this thesis, best alternative selection which is a statistical mismatch compensation method is proposed. This method aims to eliminate the high area and power cost of the conventional mismatch compensation techniques.

In order to demonstrate the proposed techniques, a low-power, 25 MS/s SAR ADC which is reconfigurable between 5-12 bits is designed. It utilizes best alternative selection method for mismatch compensation in its DAC array and comparator. Simulation results show that the analog power dissipation of the ADC can be scaled efficiently which verifies the reconfiguration concept. The designed ADC is fabricated in TSMC 65 nm CMOS. It occupies a silicon area of 998 μm X 2040 μm.

Operational tests show that the ADC functions properly. Especially, the comparator offset is reduced from19.83 mV to only 0.29 mV as a result of the compensation method.



BELMAS François
University: Grenoble INP

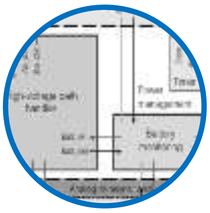
Innovating Structures for Low Power & Inductorless RF Low Noise Amplifier (LNA)

Wireless Sensor Networks (WSN) and Wireless Personal Area Network (WPAN) are expected to play a key role in our future environments. The large scale spreading of such sensors has been enabled through the strong cost optimization of modern CMOS technologies. The autonomy improvement of such sensor is however a primary concern to allow any kind of remote operation within the limitation of battery life. Those requirements of autonomy along with the context of CMOS technology development push sometimes fundamental contradictions between circuit's miniaturization and decreased power consumption.

In this work, we propose solutions to address simultaneously those autonomy-miniaturization requirements. The study presented here is focused on Low Noise Amplifiers (LNA) and more precisely on the specific case of inductorless design of LNA. Several innovative solutions has been proposed and realized in 65nm & 130nm CMOS technologies in order to highlight the pros and the cons of such design approach. The first part of this work is focused on the design of an active inductance to address the area occupation of narrow band system using inductors. We explain why such approach rises fundamental limits for radio application.

A second part details the design of an ultra-low power broadband LNA without inductors. The proposed circuits enable significant improvement in performance tradeoffs for such low power consumption in comparison with known design techniques.

PhD degrees awarded in 2013



CHRISTMANN Jean-Frédéric
University: Grenoble INP

Energy Harvesting based Power Supply Architecture and Event-driven Management for Autonomous Wireless Sensor Nodes

Wireless Sensor Networks development leverages recent progress in power consumption and in energy harvesting technologies in order to create smart sensing structures. Thanks to environmental on solar, thermal or mechanical sources, a system containing sensors and wireless communication can be powered.

This PhD works aim to study energy management within a sensing wireless node. Thanks to the use of advanced multiple power paths architecture the power management system can optimize its energy efficiency when energy is harvested. However, a precise digital control is mandatory to continuously determine the best power path between sources, loads and storages.

An integrated asynchronous controller implements an event-driven management of the power paths. This controller, implemented in QDI logic exhibits an high intrinsic robustness to environmental energy variations and ultra-low power consumption.

A power management circuit has been designed and fabricated in 180nm technology. It includes both power paths architecture and digital management innovations. Its global power consumption, close to 1µW enables the development of ultra-low-power wireless sensor nodes.

JOUBERT Antoine
University: Grenoble INP

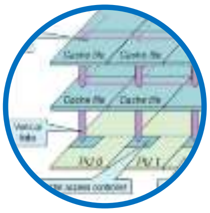
Exploring analog neuron integration in computational neuromorphic architecture from advanced towards emerging technologies

In aggressive technology nodes down to dozen of nanometers, a need of high energy efficiency has emerged. Consequently designers are currently exploring heterogeneous multi-cores architectures based on accelerators. Besides this problem, variability has also become a major issue. It is hard to maintain a specification without using an overhead in term of surface and/or power consumption. Therefore accelerators should be energy efficient and robust against fabrication defects. Neuromorphic architectures, especially spiking neural networks, address robustness and power issues by their massively parallel and hybrid computation scheme. As they are able to tackle a broad scope of applications, they are good candidates for next generation accelerators.

This PhD thesis contributed to two main aspects. Our first and foremost objectives were to specify and design a robust analog neuron for computational purposes. It was designed and simulated in a 65 nm process. Used as a mathematical operator, the neuron was afterwards integrated in two versatile neuromorphic architectures. The first circuit has been characterized and performed some basic computational operators.

The second part of this work explores the impact of emerging devices in future neuromorphic architectures. The starting point was a study of the scalability of the neuron in advanced technology nodes; this approach was then extended to several technologies such as Through-Silicon-Vias or resistive memories.

PhD degrees awarded in 2013



GUTHMULLER Eric
University: Paris VI

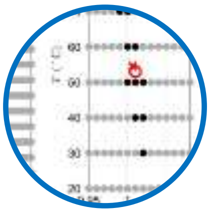
Adaptive cache architecture exploiting 3D stacking in a ManyCore context

The parallelization of processors has led to an increased need of external memory bandwidth. As the number of cores grows, it becomes difficult to embed enough memory next to processors. In this thesis, we propose a 3D cache architecture for manycore exploiting 3D stacking technologies to surpass the limitations of existing architectures. This architecture consists of a regular mesh of cache tiles interconnected by 3D networks on chip and forms a non-uniform distributed cache.

This 3D cache is reusable in a lot of contexts in order to reduce the production cost. In particular the architecture is stackable: several dies can be stacked, all dies being identical. Moreover, the 3D cache can adapt itself to the underlying topology of the processing architecture to loosen the requirements on this layer. This 3D cache also adapts itself to the needs of the application running on the processing architecture. At last, this 3D cache is tolerant to permanent faults, being able to operate in a deteriorated mode without impacting to much the performances

We have sized the 3D cache in order to reach the best energy efficiency for a reasonable hardware cost. Trials have showed that a fine-grained architecture has the best performance per cost ratio.
We have evaluated the efficiency of adaptive mechanisms implanted in the architecture and showed that these mechanisms improve its efficiency. We have also compared our architecture to the state of the art 3D memories.

Finally, in order to demonstrate the feasibility of our proposed architecture and measure the power consumption of the 3D cache, we have done its hardware implementation in a 28 nm CMOS process.



VINCENT Lionel
University: Montpellier II

Local dynamic management of variability and power consumption in MPSoC architectures

Nowadays, as embedded systems require high performance and low power, the search for the optimal efficiency of the processors, especially complex MPSoCs, has become a major challenge. Significant improvements of the energy efficiency can be expected from the reduction of the operating margins generally taken into account to ensure the circuit robustness in presence of process, voltage and temperature variations

This work is part of a low-cost architectural solution, based on local AVFS optimization technique, to reduce design margins. The development of a monitoring system of local and dynamic voltage and temperature variations using a low-cost sensor has been proposed. A first method estimates jointly voltage and temperature using statistical tests. A second one speeds up estimation of the voltage. Finally, a calibration method associated with the two previous methods has been developed.

This monitoring system has been validated on a hardware platform to demonstrate its operational nature. Taking into account the estimation of voltage and temperature values, policies to dynamically adjust the set point of the local voltage and frequency actuators have been proposed. Finally, the additional power consumption due to the integration of the components of the architecture AVFS was evaluated and compared with achievable reductions in operating margins consumption. These results showed that the AVFS solution can achieve substantial power savings compared to conventional DVFS solution.

PhD degrees awarded in 2013

BERTOLINI Clément

University: Bordeaux 1

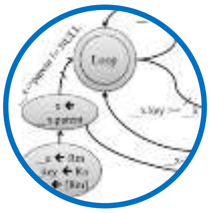
Estimation and auto-adaptative control of ageing in processing elements

The shrinking size of transistors and wires results in increased device density, increased speed and reduced power consumption. However, device reliability is reduced due to the non-ideal scaling of supply voltage. We observe two trends in semiconductor industry.

Firstly, MOS technology scaling is still continuing. The failure physics become more complex and new failures, that were negligible in old technology modes, now emerge.
Secondly, Multi-Processor SoCs (MPSoCs) offer high performance; are cheap, consumes less power than high-end processors; and are able to support a large variety of applications. As MPSoCs are now applied in most market segments (automotive, consumer, HPC, etc.), both high-performance and reliability become major concerns, even for non-safety-critical applications. MPSoC design and verification require new methodologies and CAD tools able to capture both architecture design and reliability at high abstraction level.

Two major failure mechanisms in MOS technologies are considered: hot carrier injection (HCI) and negative bias temperature instability (NBTI). The approach considers a worst case scenario that leads to conservative design. Moreover, due to new technology, margins are becoming more and more important and MPSoC performance is getting lower.

In this thesis, we propose a new methodology to simulate aging in a processor core at functional level. We propose an HCI degradation model to estimate the degradation of processor slack times according to the executed instruction. Then, we extend an existing instruction set simulator with the ability to estimate the timing degradations based on the proposed model. Finally, we validate our methodology to a processor case study. Our solution is a first step to enable architecture design exploration for MPSoC under aging constraint.



CARBON Alexandre

University: Paris 6

Hardware Acceleration for Just-In-Time Compilation in Embedded Systems

Developed since the 60s, JIT compilation is widely used for 15 years. This is the consequence of two main phenomena: the increasing dynamism of applications and the increasing demand concerning virtualization.

The transfer of these issues to the embedded domain leads to experience JIT compilation on small and sparse resources. However, the management of JIT compilation algorithms' complexity and irregularity on small resources (in-order processors, limited speculation, limited memory hierarchies) leads to important scaling-down problems in terms of performance. As a consequence, JIT compilation solutions are less attractive in this domain.
While several software optimizations have been already proposed in the literature, we propose in this thesis the development of hardware accelerations coupled to the processor in charge of the JIT compilation. The final aim is to propose a more efficient solution in terms of performance with respect to embedded constraints.

Based on the LLVM framework compiler (LLC), our experiments highlight two critical points in terms of performance: the associative array and dynamic memory allocation management and the instruction graph handling for instructions to compile and optimize. Two accelerators have been proposed in this way. Concerning the management of associative arrays, we obtain gains up to 25 % on LLC with an area overhead under 1.4 % of the associated processor.

PhD degrees awarded in 2013

ut: V, N,
1: $g_b \leftarrow$ null
2: for $i = 1$ to Max
3: $g_c \leftarrow$ construc
4: $g \leftarrow$ local_sea
5: update best as
6: end for
ut: best assi

STAN Oana
University: UTC, Compiègne

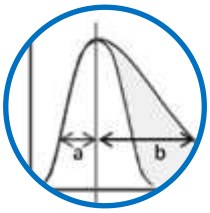
Placement of tasks under uncertainty on massively multicore architectures

This thesis is devote to the study of combinatorial optimization problems related to massively parallel embedded architectures when taking into account uncertain data (e.g. execution time). Our focus is on chance constrained programs with the objective of finding the best solution which is feasible with a preset probability guarantee.

A qualitative analysis of the uncertain data we have to treat (dependent random variables, multimodal, multidimensional) has lead us to design a non-parametric method, the so-called ``robust binomial approach'', valid whatever the joint distribution and which is based on robust optimization and statistical hypothesis testing. We also propose a methodology for adapting approximate algorithms for solving stochastic problems by integrating the robust binomial approach when verifying for solution feasibility. The practical relevance of our approach is validated through two problems arising in the compilation of dataflow application for manycore platforms.

The first problem treats the stochastic partitioning of networks of processes on a fixed set of nodes, by taking into account the load of each node and the uncertainty affecting the weight of the processes. For finding stochastic solutions, a semi-greedy iterative algorithm has been proposed which allowed measuring the robustness and cost of the solutions with regard to those for the deterministic version of the problem.

The second problem consists in studying the global placement and routing of dataflow applications on a clusterized architecture. The purpose being to place the processes on clusters such that it exists a feasible routing, a GRASP heuristic has been conceived first for the deterministic case and afterwards extended for the chance constrained variant of the problem.



ELSAHMARANY Lola
University : Clermont-Ferrand 2

Methods for improving wire diagnosis using reflectometry

The use of electric cables has been significantly increasing over the last decades. However,the reliability of these systems is partially based on the reliability of electrical networks. A significant number of failures and malfunctions of these systems come from faults in wired links and not from electrical devices. Therefore, the knowledge of the state of wire networks and particularly the detection of their faults is important.
Several methods have been developed to test the status of cables. Among them, reflectometry methods are widely used and easily embeddable. Improvements in measurement and processing are necessary to overcome the limitations of these methods.

In this thesis, three new methods for wire diagnosis have been studied and developed to improve and ease the detection and location of soft wire faults:
– The first method, called "adaptive correlation", provides a new algorithm to compensate signal’s dispersion. It improves faults location and the detection of singularities on cables.
– The second method, called TRR (Time Reversal Reflectometry), is based on the principle of reflectometry and time reversal. It allows the characterization of aging of electrical cables.
– The third method, called RART (Reflectrometry combined with a time reversal process), is also based on the principle of reflectometry and time reversal. It improves the detection of electrical faults related to degradation of insulation.

This research illustrates the efficiency and applicability of the proposed methods. It also demonstrates the potential of the proposed methods to improve safety in operation of electrical systems whether in transport, construction, or even communication networks.

Notes

Notes

Greetings

Editorial Committee

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

Readers

Thierry Collette

Graphic Art

Valérie Lassablière

Contacts

Thierry Collette
Head of « Architecture, IC Design & Embedded software » Division
thierry.collette@cea.fr

Ahmed Jerraya
Head of the Design Center Initiative
ahmed.jerraya@cea.fr

Cyril Condemine
Smart Sensor Integration Program Manager
cyril.condemine@cea.fr

Michel Durr
Analog, Imaging & wireless IC Program Manager
michel.durr@cea.fr

Benjamin Lucas-Leclin
Embedded Systems Program Manager
benjamin.lucas-leclin@cea.fr

Marc Duranton
Research Director, European projects Coordinator
marc.duranton@cea.fr

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

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

Paul Grève
Business Development
paul.greve@cea.fr

CEA - Leti
CEA Grenoble
17, rue des Martyrs
F-38054 GRENOBLE Cedex 9
Tel. (+33) 4 38 78 37 29
www.leti.fr

CEA - List
CEA Saclay, NanoInnov
F-91191 Gif sur yvette
Tel. (+33) 1 69 08 49 67
www.list.cea.fr