Power mitigation of a heterogeneous multicore architecture on FPGA/ASIC by DFS/DVFS techniques

Sajjad Nouri [a],[*], Davide Rossi [b], Jari Nurmi [a]

[a] *Laboratory of Electronics and Communications Engineering, Tampere University of Technology, Tampere, Finland*
[b] *Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI), University of Bologna, Bologna, Italy*

ABSTRACT

This article presents an integrated self-aware computing model in a Heterogeneous Multicore Architecture (HMA) to mitigate the power dissipation of an Orthogonal Frequency-Division Multiplexing (OFDM) receiver. The proposed platform consists of template-based Coarse-Grained Reconfigurable Array (CGRA) devices connected through a Network-on-Chip (NoC) around a few Reduced Instruction-Set Computing (RISC) cores. The self-aware computing model exploits Feedback Control System (FCS) which constantly monitors the execution-time of each core and dynamically scales the operating frequency of each node of the NoC depending on the worst execution-time. Therefore, the performance of the overall system is equalized towards a desired level besides mitigating the power dissipation. Measurement results obtained from Field-Programmable Gate Array (FPGA) synthesis show up to 20.2% dynamic power dissipation and 16.8% total power dissipation savings. Since FCS technique can be employed for scaling the frequency and the voltage and on the other hand, voltage supply cannot be scaled on the FPGA-based prototyped platform, the implementation is also estimated in 28nm Ultra-Thin Body and Buried oxide (UTBB) Fully-Depleted Silicon-On-Insulator (FD-SOI) Application-Specific Integrated Circuit (ASIC) technology to scale voltage in addition to frequency and get more benefits in terms of dynamic power dissipation reduction. Subsequent to synthesizing the whole platform on ASIC and scaling the voltage and frequency simultaneously as a Dynamic Voltage and Frequency Scaling (DVFS) method, significant dynamic power dissipation savings by 5.97X against Dynamic Frequency Scaling (DFS) method were obtained.

© 2018 Published by Elsevier B.V.
This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

In 1965, Moore's Law claimed that the number of transistors that can be integrated on a single chip doubles every one to two years [1]. After that in 1974, Dennard showed that the power dissipation density can be kept constant by scaling the CMOS devices while voltage and current should be proportional to the linear dimensions of a transistor [2]. However, this trend is not continuing anymore since Dennard ignored the baseline of power per transistor established by leakage current in modern technologies, resulting to enormous heat dissipation. Then RAW microprocessor architecture as a multicore has been introduced in order to solve the problem of heat dissipation and power density by exploiting several processors operating in parallel at a lower frequency in-

stead of a single core operating at very high frequency [3]. Although introducing RAW microprocessor was a big step for dealing with the mentioned issue, the experimental work explained in [4] showed that only 7% of a 300 mm^2 chip can be operated at full frequency under a power budget of 80W which put an end to multicore scaling. It also stated that on a single chip, all cores cannot be clocked at their maximum operating frequency for a given Thermal Design Power (TDP) constraint which forced a large fraction of chip to be to be switched-off (dark) or to be operated at very low frequency (dim). Then, four different approaches as *Four Horsemen* (Shrink, Dim, Specialized and Deus Ex Machina [5]) have been suggested in order to deal with the issue of "utilization wall" that causes the Dark Silicon problem [4]. Among them, *Specialized Horseman* proposed domain-specific accelerators and in particular Coarse-Grained Reconfigurable Arrays (CGRAs) to be instantiated to avoid the dark and dim parts of the chip [5]. CGRAs are operated at a very low frequency and can be reconfigured at runtime for multiple applications and yield tremendous performance

* Corresponding author.
E-mail addresses: sajjad.nouri@tut.fi (S. Nouri), davide.rossi@unibo.it (D. Rossi), jari.nurmi@tut.fi (J. Nurmi).

improvements exploiting computational units working in parallel [10]. Thus, the dark part of the chip can be employed for performing massively-parallel workloads of critical-priority applications [5]. Consequently, the execution time of those kernels would be reduced significantly in addition to operating at a very low frequency.

Despite all the benefits can be obtained from CGRAs, they have potentially high power dissipation. Power consumption can be mitigated by employing various techniques such as Dynamic Voltage and Frequency Scaling (DVFS) and clock gating [8]. There are several case-studies in the literature, presented in the next section, which show the effects of DVFS method for mitigating the power and maximizing the performance. The thermal problems can be solved by employing DVFS which keeps the temperature under a given threshold [6]. According to the level of granularity, DVFS schemes can be varied from fine-grained to coarse-grained which are suitable for modifying the frequency of each resource separately (better energy efficiency) or scaling the performance of the whole platform (maximum performance), respectively [7]. With the rising complexity of extreme-scale computer systems and increasing the need of maximum performance besides minimum energy consumption, SElf-awarE Computing (SEEC) models have been proposed in order to create self-adaptive computing systems with the ability of changeable behavior according to the performance requirements [8]. In this context, Feedback Control System (FCS) is an advanced technique to reduce the dynamic power dissipation by constantly monitoring the execution time and recognizing the worst-case one and then scaling the frequency and the voltage to meet the desired performance level. These systems can also exploit the inertia of heat dissipation to boost performance of some resources which can be useful for execute some urgent task, but only for short times only to avoid overheating [25].

This experimental work presents the power mitigation of the Heterogeneous Accelerator-Rich Platform (HARP) [9] platform to be used as an Orthogonal Frequency Division Multiplexing (OFDM) receiver by applying Dynamic Frequency Scaling (DFS) method on Field-Programmable Gate Array (FPGA) and also on Application-Specific Integrated Circuit (ASIC) technology. The designed platform, modified to be used as an OFDM receiver, is composed of seven nodes, three Reduced Instruction-Set Computing (RISC) processors and four specialized template-based CGRAs in order to perform a fine mixture of parallel and serial tasks provided by OFDM receiver. It should be mentioned that although one RISC core is also sufficient for performing this task, however due to the data dependency between the nodes to perform an OFDM receiver completely as well as increasing the level of parallelism, three RISC cores are employed. The RISC processors are responsible for continuously monitoring the execution time of each computing-node, calculating the worst-case execution time and then upgrading or downgrading the clock frequencies and supply voltage of the slave nodes to minimize the overall power dissipation. Furthermore, the power overhead of these controlling cores is also taking into account when the energy improvement of FCS is computed. For this purpose, the FCS technique is implemented in the processor software of three RISC cores to perform dynamic frequency scaling by identifying the worst-case execution time and then tune the operating voltage and frequency of the computing-nodes to meet the performance thresholds. In order to evaluate the benefits of voltage scaling on top of the FCS technique, we have also synthesized the HARP system in 28nm Ultra-Thin Body and Buried oxide (UTBB) Fully-Depleted Silicon-On-Insulator (FD-SOI) ASIC technology as an extension version of [30] and estimated the power consumption of the nodes at different supply voltages, starting from the nominal voltage supply of the technology (1V) down to the deep near threshold regime (0.5V). Therefore, power saving can be guaranteed as a vital issue for telecommunication systems while dark and dim part of the chip can remain operable because of overall heat dissipation reduction. The proposed approach reduces the dynamic power consumption on FGPA by 20.24% and by 97.61% on ASIC when exploiting DVFS.

The rest of the paper is organized as follows. Section II presents a brief survey of already existing related work about DVFS method in multicore platforms. Section III describes shortly the architecture of HARP platform with its overall functionality, the architecture of integrated template-based CGRAs, the internal structure of the Network-on-Chip (NoC) nodes and the execution flow. In Section IV, applying the FCS technique on OFDM receiver test case for mitigating the power and equalizing the performance of the overall multicore platform on FPGA and ASIC is provided. Section V presents the achieved results in terms of power saving as well as measurements and estimations of different performance metrics subsequent to applying the FCS technique with frequency and voltage scaling. Finally, in Section VI, conclusions are discussed.

2. Related work

HARP is designed in order to maximize the number of computational resources for accelerating many specific computationally intensive algorithms besides investigating the issues related to Dark Silicon. In addition to HARP, other state-of-the-art homogeneous and heterogeneous computing platforms have been introduced so far such as Platform 2012 ([11,12]) and MORPHEUS ([13,14]) which is one of the most promising heterogeneous platforms. MORPHEUS has been designed by Rossi et al. [15] as a heterogeneous digital signal processor for dynamically reconfigurable computing based on a 64-bit NoC. The platform has been developed by combining a Fine-grained embedded FPGA, a Mid-grained configurable processor and a CGRA for exploiting DFS to mitigate the power consumption. The supervisor node is an ARM 926EJ-S RISC processor which monitors communication, synchronization and reconfiguration mechanisms. This platform has been fabricated in 90-nm technology with the size of 110 mm^2 delivering 120 GOPS on a video surveillance motion detection application with the power dissipation of 1.45 W and peak power consumption of 2.5 W. In another case-study [16], Fulmine, a 65 nm System-on-Chip (SoC) based on a tightly coupled multicore cluster supported with specialized blocks for computationally intensive tasks has been developed to be used as the emerging class of smart secure near-sensor data analytics for Internet-of-Things (IoT) end-nodes without voltage scaling. The proposed multicore platform achieved the power consumption of 20 mW on average at 0.8V, up to 25 MIPS/mW in software, 315 GOPS/W, low energy, potentially high speed and low-effort data exchange among processing engines. As a general comparison among the state-of-the-art homogeneous and heterogeneous computing platforms, Fulmine is heavily specialized for IoT application and does not support general purpose, MORPHEUS does not support template-based CGRAs hence might be worst in terms of performance, while HARP provides a better trade-off in terms of general purpose ability, performance and scalability since the template-based CGRA can be specialized for different application domains. Moreover, in the case of HARP, computational resources can be allocated at design time such that a core does not over or under perform relative to the overall execution time frame.

Many authors have tried to mitigate the power dissipation of different homogeneous/heterogeneous multicore platforms by using DVFS techniques and SEEC models. In [10], the authors have applied software-defined FCS on a heterogeneous platform and equalized its performance which resulted in mitigation of the overall dynamic power dissipation by 20.7%. In another case study [17], the thermal problems of 3D multicore processors, caused by high power density because of the stacking of multiple layers ver-

tically, are investigated by employing an adaptive DFS technique which resulted in reduction of the peak temperature by up to 10.35 °C. In [18], the energy level of shared resources such as NoC and Last-Level Caches (LLCs) in multicore processor designs is reduced by 56% by applying a DVFS method. In [19], energy aware task parallelism has been presented for CGRAs which relies on resource allocation graphs, autonomous parallelism, voltage and frequency selection algorithms. They achieved considerable reduction in energy, power and cofiguration memory requirements of up to 36%, 28% and 36%, respectively in comparison with three different state-of-the-art DVFS algorithms. In another case-study [20], proportional integral derivative controller technique as a DVFS technique was proposed for manycore systems to keep the operating temperature within the thermal design power bounds and therefore, mitigate the power consumption based on the certain power limit. They proved the effectiveness of their method by enhancing the system throughput up to 43%. In [21], authors achieved 8.4X energy saving in a self-aware processor with the ability of self-adaption by monitoring the energy consumption. In [22], authors concluded that self-aware systems can play an important role for the IoT systems by adding the following features: enabling complex high-volume IoT architectures, performing configuration and adaption at run-time, enabling safety-critical application by adding planning and modeling to the IoT system's infrastructure and etc. In another case-study [23], the authors employed DVFS technique to a homogeneous multi-processor architecture and achieved the improvement of the overall system performance by a factor of 3 compared to clock gating and also considerable energy saving. In a case-study performed by Rossi et al. [24], a software controllable self-aware architecture exploiting Body Biasing (BB) has been implemented in 28-nm ultra thin body and box fully depleted silicon on insulator technology for compensation of parameters, operational voltage and temperature and for implementation of low-power modes in near-threshold processors. BB is an advanced technique in which a voltage will be applied to the body contact of CMOS transistors and accordingly, the effective transistor threshold voltage can be shifted and also the leakage power consumption can be reduced. In this study, the wide range forward BB (FBB) and reverse BB (RBB) were employed for reducing the design time margins and introducing a low-power mode and state-retentive sleep mode. According to the conducted experiment results in [24], design margins reduced enormously while the energy efficiency of the processor improved by 32% with a hardware cost of less than 1% and a runtime cost for software control of less than 0.01%. Furthermore, 24x area reduction for the compensation loop and 21.2x better efficiency were achieved in [24] compared to their previous design.

In this paper, the proposed FCS technique with the high-level of complexity due to applying on three RISC cores simultaneously, reduces the overall power consumption (as a vital issue for telecommunication systems and IoT purposes) significantly on FGPA and ASIC when exploiting DFS/DVFS methods.

3. The heterogeneous multicore platform

The HARP template, as is depicted in Fig. 1, consists of nine nodes arranged in a topology of three rows and three columns over a NoC. The central node which is responsible for General-Purpose Processing (GPP) and overall platform supervision is always integrated with a RISC core (called COFFEE [26]) while the rest of the nodes can be employed to be a RISC core or instances of the template-based CGRAs as coprocessors for performing computationally intensive tasks. According to the application requirements, nodes can exchange data between each other at run-time in the case of data dependency or even work independently and simultaneously.
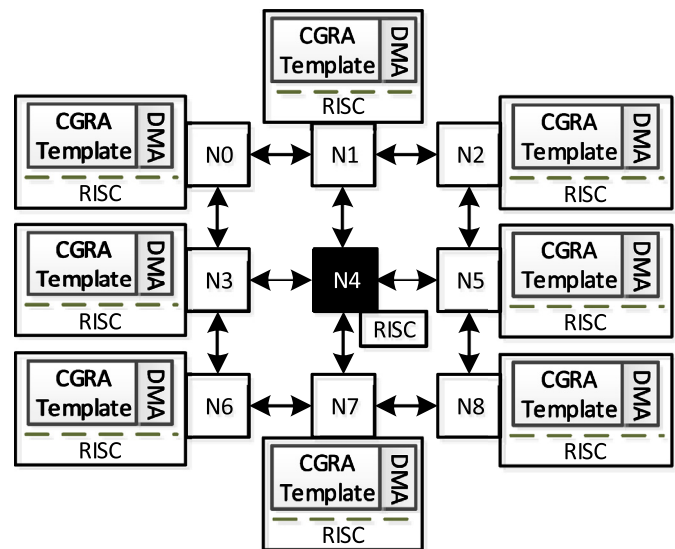

Fig. 1. An overview of the HARP architecture applying a processor/coprocessor model [29].

The template-based CGRAs integrated in the HARP template (called CREMA, AVATAR [27]), shown in Fig. 2, vary in terms of their sizes based on the application requirements while their architectural features are almost the same. Template-based CGRAs are equipped with the arrays of Processing Elements (PEs) while the number of rows and columns of PEs can be scaled-up/down based on the proposed applications and their algebraic expressions in order to be performed efficiently. The functionality of each PE and also interconnection between PEs in a point-to-point fashion with multiple routing possibilities can be specified at design time by the designer. The internal structure of each PE is depicted in Fig. 3. Every PE receives two input operands and executes a 32-bit integer or a floating-point operation (IEEE-754 format). A PE is composed of a Look-Up Table (LUT), adder, multiplier, shifter, immediate register and floating-point logic. The blocks which are shown with the dashed borders are selectivity instantiated according to the processing requirements of an algorithm's algebraic expression at design-time while the two input operand registers are always instantiated. At runtime, the functionality and routing are controlled by reconfiguration. The data can be interleaved between data local memories and the PEs by using the I/O buffers which are made of sixteen 16 or 32 × 1 multiplexers and 16 or 32 32-bit registers for CREMA and AVATAR, respectively. Each PE has interconnections with neighboring PEs in a point-to-point fashion with the following routing possibilities, i.e., local, interleaved and global. In order to perform an application, a number of configuration contexts, designed at the system design-time and enabled at run-time, may required. A context is the pattern of interconnections among all PEs and the set of operations to be performed by each PE at any clock cycle. According to the execution flow of an application, the contexts can be switched at run-time by deploying the configuration words, consisting of an address and operation field to determine the task of each PE and its destination address, respectively. They are stored during the system startup time at configuration memory of PEs by the Direct Memory Access (DMA) device [28]. The overall control flow of the designed template-based CGRAs, performed by COFFEE RISC core, is programmable in C language. Once an application-specific accelerator is designed, it can be integrated with one of the existing network nodes.

As it can be observed from Fig. 1, nodes are connected to each other in a point-to-point fashion. The slave nodes integrated with the template-based CGRA can exchange data between their local
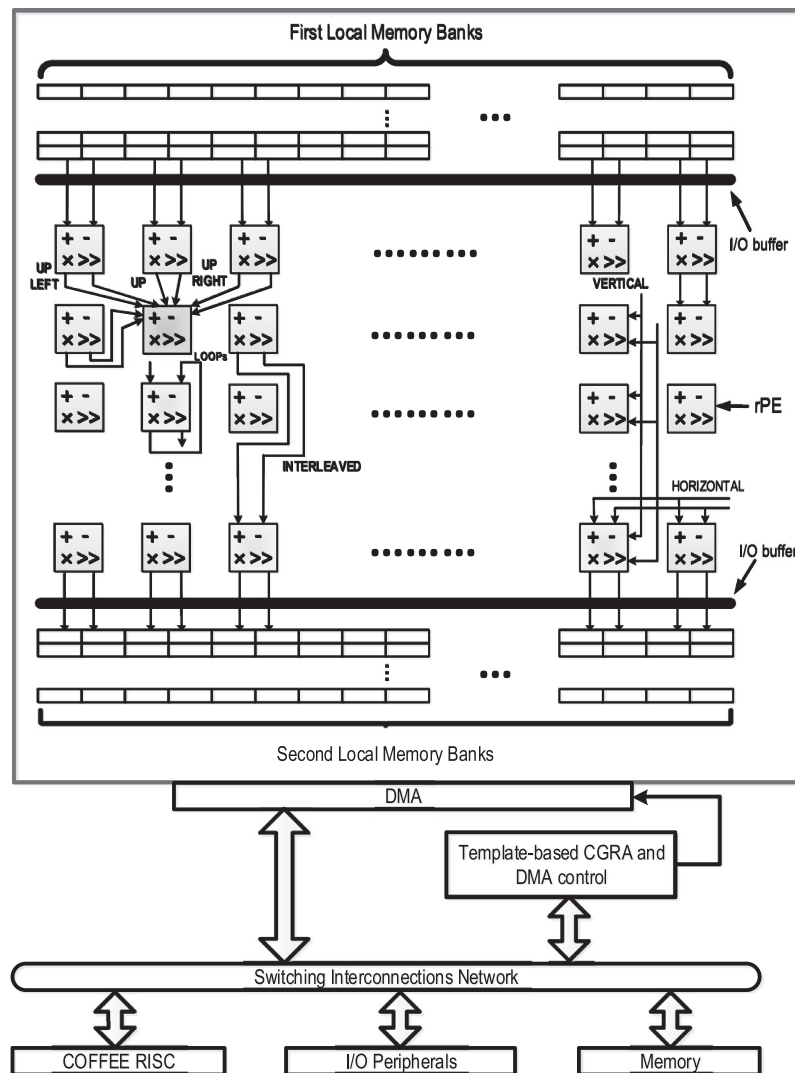
Fig. 2. The architecture of scalable template-based CGRA [29].

memories directly or with the help of RISC cores. Each node of the NoC has one master and two slave interfaces. The master one, integrated with the RISC core, can be employed for writing to the network and transferring data within a node while the slave interfaces are used for controlling the clock frequency and supply voltage selection, and integrating the data memory. The central node is integrated with the RISC core as a supervisor node for transferring data in the form of packets among its own data memory and data memories of the slave nodes. The target devices for transferring the data packet gets selected by using the switches integrated in each node according to the information in the routing field of the transported packet which can point to the data memory of the respective node or the DMA slave. Moreover, in the case of data dependency between the nodes, the supervisor node is responsible for establishing synchronization for data transfer between two different nodes by using an allocated shared memory space. The software/hardware co-design flow for template-based CGRAs integrated over HARP can be listed as follows:

- Defining the functionalities of the PEs and routing among them by using a Graphical User Interface (GUI) tool;
- Generating the configuration files for mapping and run-time reconfiguration;

- Loading the configuration data in the template-based CGRAs at the system start-up time by using DMA device which can be used for switching the contexts and performing reconfiguration;
- Loading the data to be processed into one of the local memories of the template-based CGRAs by using the DMA device monitored by the host RISC core;
- Enabling a context for configuring the functionalities of the PEs and the routing between them;
- Processing the data over the PE array;
- Switching the context in order to reconfigure the template-based CGRAs for performing the new task;
- As required, loading the new set of data and iterating the above-mentioned steps;
- Transferring back the results from the local memory of a CGRA node to the data memory of its host RISC core for further processing;
- Iterating the above phases until the algorithm completes its execution.

More information about the execution flow of performing an particular application by the use of template-based CGRAs as well as the application mapping on the HARP platform and the internal structure of the NoC can be found in [29].
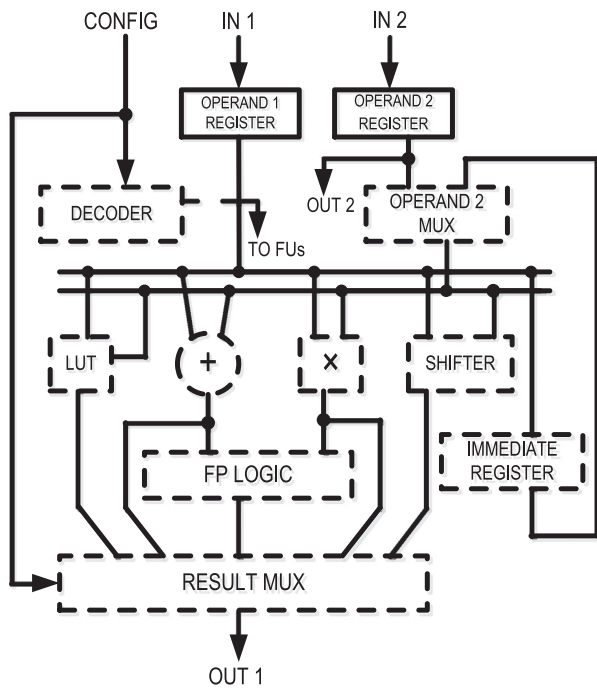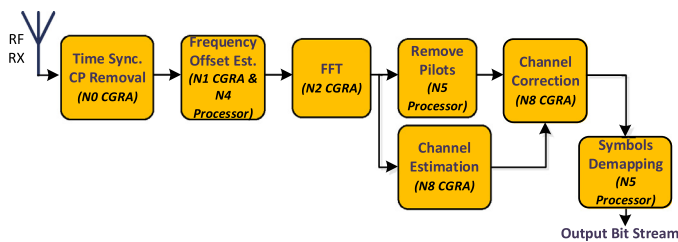
Fig. 3. PE Core Template.

Fig. 4. A simplified block diagram of an OFDM receiver.

4. Equalization of the OFDM receiver performance by frequency and voltage scaling

Prior to discussing about the equalization of the OFDM receiver performance by using DFS and DVFS methods, let us have a brief explanation about the design and implementation of an OFDM receiver blocks on HARP as well as the reasons behind selecting OFDM application as a test-case.

4.1. Design and implementation of an OFDM receiver blocks on HARP

OFDM is an important data transmission scheme in Software Defined Radio (SDR) technology because of providing high data rates and it is also a candidate to be employed for 5G wireless systems. OFDM receiver blocks as a critical part due to retrieving the data after the noisy channel, are composed of most computationally intensive and time-consuming tasks such as Fast Fourier Transform (FFT), Correlation, Convolution and Complex Matrix-Vector Multiplication (MVM) which require parallel processing, shown in Fig. 4. Furthermore, there are some tasks such as Frequency Offset Estimation which require the GPP of some serial in nature algorithms such as Taylor series and CORDIC algorithms. Such a fine mixture of parallel and serial algorithms makes the OFDM receiver as a valuable test-case in order to evaluate almost all the design features and technical capabilities of the HARP template as well as to identify potential architectural fallacies and pitfalls.

As it is depicted in Fig. 4, the parallel algorithms are implemented by crafting template-based CGRAs while the general-purpose tasks such as Frequency Offset Estimation are executed by using both template-based CGRA and RISC processor. Subsequent to mapping the kernels on the designed CGRA accelerators, they can be integrated over HARP in such a way that both master and slave nodes can exchange data with each other (Fig. 5). According to Fig. 4, the order of exchanging the data in Fig. 5 for performing the OFDM receiver blocks should be initiated by N0 CGRA node (belonging to Time Synchronization block). After passing through N1 and N2 which are associated with Frequency Offset Estimation and FFT, respectively, ending up with N8 CGRA node (at Channel Estimation block). The detail of design and implementation of each block of an OFDM receiver which led to scale CGRAs to different dimensions can be found in [29].

4.2. Performance equalization by DFS/DVFS methods

The overall architecture of HARP, modified for performing an OFDM receiver functionality is depicted in Fig. 5. Each of three RISC cores acts as an controller and observer for monitoring the performance of their associated CGRA nodes and transferring the configuration stream and data to be processed in a way that node N3 RISC is responsible for N0 CGRA while N4 RISC is responsible for N1 CGRA and N5 RISC is responsible for N2 and N8 CGRAs. Since there are multiple CGRA nodes with different dimensions in the platform those are supposed to exchange data with each other, running a core faster than the other is not desirable. Subsequent to transferring the configuration stream by three RISC cores in parallel, the data will be loaded into the local memories of N0 in order to implement Time Synchronization block. Then the number of clock cycles required for the execution of the Time Synchronization block counted by using a special counter of the RISC processor. The counted number of clock cycles for complete execution of Time Synchronization block will be transfered to the N4 RISC core and stored at reserved locations in its data memory for seeking the worst-case execution time. Total clock cycles required for performing each block of OFDM receiver contains the following stages: data memory to data memory, data memory to local memory of the CGRA and vice versa and also the CGRA execution time. Since each computing-node has to wait until the other one completes its execution because of the data dependency between the receiver blocks, all three RISC cores establish synchronization between each other by writing 'read' and 'write' flags in their shared memory location.

The CGRA accelerator blocks of the platform form a software-defined macro-pipeline. By completion the process of Time Synchronization block, an acknowledgment will be sent by DMA's master from N3 to N4 which gives the permission to N4 for starting the process of Frequency Offset Estimation block. Other CGRAs will also perform their tasks with the same procedure and at the end of the first iteration, the data related to the counted number of clock cycles of four worker-nodes will be retrieved in order to recognize the stored most time consuming CGRA (worst-case). The total number of clock cycles related to the nodes N0, N1, N2 and N8 are equal to 9985, 18039, 1931 and 1685 CC, respectively [29]. Based on the achieved results, node N5 CGRA, which contains the Frequency Offset Estimation block, is the most time consuming one and can be selected as a worst-case candidate by master node N4 RISC core. Therefore, N4 as a supervisor node will notify the other two RISC cores about the selected worst-case candidate. From the second round of iteration, nodes N3, N4 and N5 RISC cores will tune the operating frequency of the CGRA cores belonging to them in order to approach the defined equalization region. The FCS technique has been implemented completely in RISC software to perform dynamic frequency scaling by reducing
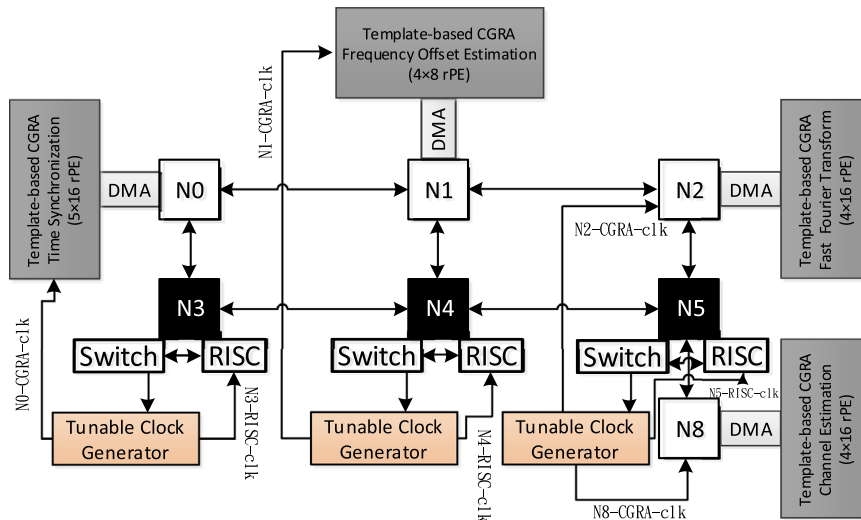
Fig. 5. A simplified overview of the HARP architecture with three RISCs and four template-based CGRAs in a processor/coprocessor model [30].
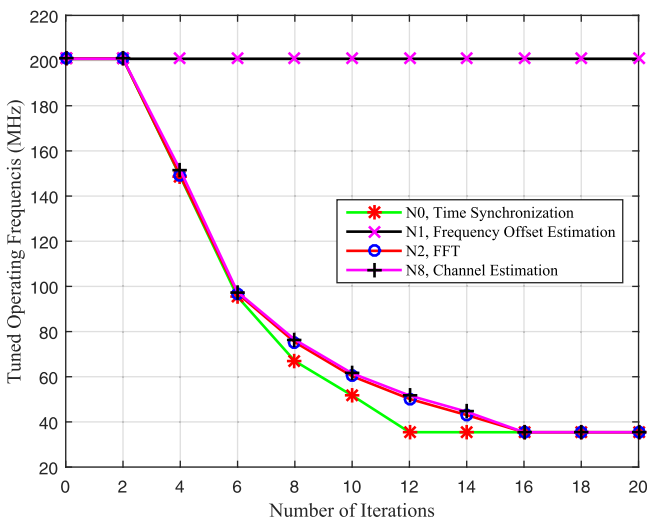

Fig. 6. Tuning the operating frequencis in the range of ≈ 35.0–200.0 MHz on FPGA prototype [30].
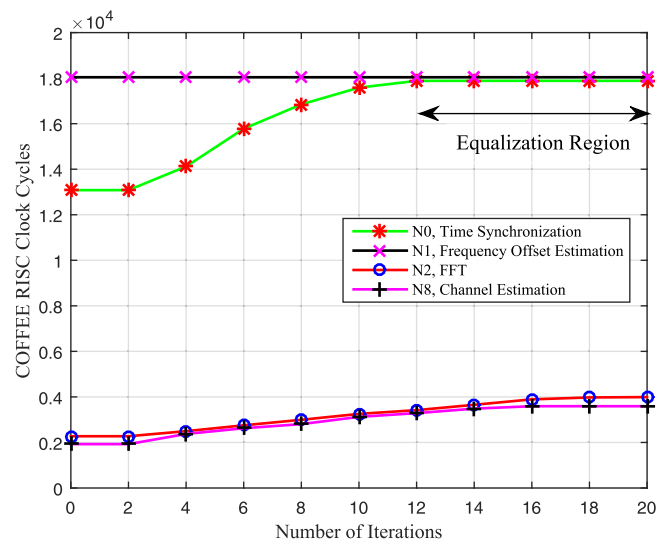

Fig. 7. Performance Equalization of the CGRAs based on the Worst-Case Execution Time on FPGA prototype [30].

each core operating frequency to approach the performance goal which can be identified by comparing the counted clock cycles of the CGRAs and selecting the worst-case execution-time. The clock frequencies of the CGRAs can be updated through a module emulating a DVFS Power Management Unit (PMU). This PMU includes a 32-bit general-purpose register, defined as allocated 4-bit field (16 bits in total for four cores) for each one when in DVFS mode (on ASIC), the corresponding supply voltage supporting the clock frequency is also selected. When the frequency is updated the PMU clock gates the part of the systems where the frequency or voltage is changed. This guarantees that the subsystem is not operating during the transitions of frequency and voltage, and that operation is safely restored once the voltage/frequency are stable.

Each RISC core can tune the clock frequency of its associated CGRA by using this 4-bit field between 16 different frequencies in the range of 35.0–200.0 MHz on the FPGA prototype (shown in Fig. 6). Meanwhile, the operating frequency of the three RISC cores is maintained fixed at 100.0 MHz and is set as a reference of all measurements. At each iteration, FCS software adds or subtracts a 4-bit field by "0001" in order to update the current operating frequency of the CGRA. It has to be mentioned that during the itera-

tions, the clock frequency of the RISC cores will remain the same. As it can be observed from Fig. 6, the clock frequency of all the CGRA cores operate at the maximum possible operating frequencies during the system start-up time, near to 200.0 MHz (achieved after placement and routing, to be explained in detail in the next section). During the first iteration, the execution time of the CGRA cores and accordingly, the worst-case candidate can be identified. From the second iteration onwards, FCS will automatically update the target and start to tune the operating frequency of other CGRA nodes for twenty iterations in order to approach the equalization region. The performance equalization of the CGRA cores is also depicted in Fig. 7. It can be seen that by reaching the 12th iteration, N0 successfully approached the goal while running at 35.0 MHz operating frequency. However, for the other two computing-nodes (N2 and N8), because of the smaller workload, lower computation complexity algorithm and relatively shorter execution time, their performance could not degraded to the equalization region even with running at 35.0 MHz operating frequency during the last iterations.

Subsequent to synthesizing on FPGA, the same procedure is performed by synthesizing the modified HARP template on ASIC

Table 1
Resource Utilization summary for Stratix-V FPGA device [30].

Node	ALMs	Registers	Block Memory Bits	(32-bit Multipliers) DSPs
N0	22,729	8,612	2,633,472	(20) 40
N1	8,255	7,855	2,364,672	(16) 32
N2	24,149	11,326	2,633,472	(28) 56
N3	5,590	5,648	3,145,728	(6) 12
N4	5,646	5,668	4,194,304	(6) 12
N5	5,601	5,700	3,145,728	(6) 12
N8	25,974	15,975	2,633,144	(33) 66
NoC	2,533	4,073	-	-
	100,477	66,823	20,750,520	(115) 230
Total	63%	11%	53%	90%

Table 2
Dynamic power dissipation of each node and the NoC before/after applying FCS on FPGA prototype [30].

Node	Dynamic Power FCS Inactive [29] [mW]	Dynamic Power FCS Active [mW]	Gain %
N0	414.35	284.43	31.35
N1	272.23	272.83	≃ 0.0
N2	526.07	391.23	25.63
N3	114.47	104.13	9.03
N4	113.82	114.61	≃ 0.0
N5	114.52	105.07	8.25
N8	448.01	344.61	23.07
NoC Integration	10.10	14.34	-
Logic	609.07	461.35	24.25
Total	2623.72	2092.6	20.24

technology and dynamically scaling both frequency and the voltage within the range of 55.0–500.0 MHz and 0.5-1V, respectively. In this case, the dynamic power dissipation of the nodes is estimated in several operating conditions, first just DFS at 1V and then DVFS, explained in detail in the next section. Furthermore, since the comparison among power mitigation of the platform on FPGA and ASIC is not a completely fair, as the clock frequencies are different, we reduced the maximum operating frequency of ASIC from 500.0 MHz to 200.0 MHz which is the same as FPGA one.

5. Measurements and estimations

The overall HARP platform with the applied FCS was first synthesized on a Stratix-V FPGA device (5SGXEA4H1F35C1) for prototyping the concept and then in 28nm UTBB FD-SOI ASIC technology for estimating the added benefits of including also voltage scaling.

5.1. FPGA Evaluation

We will first take a look at the FPGA prototype results here. Node-by-node breakdown of resource utilization on FPGA is provided in Table 1 in terms of the following metrics: Adaptive Logic Modules (ALMs), Registers, Memory Bits and DSP elements. It has to be mentioned that the number of employed 18-bit DSP resources depends on the number of multiplications performed by the template-based CGRAs where for each 32-bit multiplier, two 18-bit DSP elements are required. More details about tailoring the template-based CGRAs for performing OFDM receiver algorithms as well as instantiating 32-bit multipliers by using PEs on CGRAs based on their algebraic expressions are explained in [29]. As it can be observed from Table 1, the logic utilization increase due to applying FCS technique is just around 1% compared to [29], which is almost negligible. Subsequent to synthesizing the overall platform on Stratix-V FPGA device by using Quartus II 15.0, two timing models were selected in order to measure the operating frequencies after placement and routing. The timing models used by the Quartus II software could cover worst-case voltage to the minimum and maximum supported Vdd operating conditions for Slow 900mV 85°C and Fast 900mV 0°C, respectively. By using these timing models, the timing of FPGA can be verified without the need to implement physical simulation. In this regard, the maximum achieved operating frequencies for slow timing model at an operation voltage of 900 mV are equal to 163.61 MHz and 188.29 MHz at temperatures of 85°C and 0°C, respectively. In the case of fast timing model (900 mV), the maximum operating frequencies are equal to 246.12 MHz at 0°C and 223.51 MHz at 85°C. First of all, the clock frequency of 170.0 MHz has been used for running the simulations using the ModelSim simulator and then it is increased up to 200.0 MHz as the average of the achieved operating frequencies without any timing error on the particular FPGA instance used.

The platform's power dissipation has been estimated based on post placement and routing (post P&R) information, not functional simulations, using PowerPlay Power Analyzer Tool of Quartus II 15.0 at an ambient temperature of 25°C. The power analyzer provided us a "HIGH" confidence metric in our power dissipation estimations acquired from simulating the gate-level netlist. The power estimations are performed in two cases: inactive state of FCS with the fixed operating frequency of 200.0 MHz and active state of FCS with the tunable operating frequency after every iteration started from 200.0 MHz. Node-by-node breakdown of dynamic power dissipation for the both cases is shown in Table 2. Nodes N0, N1, N2 and N8 are belong to Time Synchronization, Frequency Offset Estimation, FFT and Channel Estimation, respectively while nodes N3, N4 and N5 are belong to General Purpose Processing, Synchronization and Control. In the case of inactive state of FCS, the tool estimated 1243.84 mW, 2623.72 mW, 27.37 mW and 3894.93 mW as static, dynamic, I/O and total power dissipation, respectively. Moreover, it can be found out that by scaling up the size of CGRA (N1 compared to N0, N2 and N5, Fig. 5), dynamic power dissipation will be increased in the range of 1.5X-2X. Once the FCS starts to tune the operating frequency of the cores based on the worst-case execution time, the dynamic power dissipation will also start to be reduced, up to 20.2% for the total dynamic power in this case study. In the case of FCS active, the estimated static, dynamic, I/O and total power dissipation for the overall platform showed the value of 1121.19 mW, 2092.6 mW, 27.37 mW and 3239.2 mW, respectively. Compared to the inactive state of FCS, the total power dissipation of the platform is reduced by 16.8%. Although it is proved that the FCS will decrease the instantaneous dynamic power dissipation and therefore the heat dissipation and the dark/dim part of the chip, the energy consumption on the FPGA prototype for completing the operations will remain approximately the same, as the decrease of clock frequency will increase the active time proportionally.

5.2. ASIC Evaluation

As the next step, the overall platform is synthesized on 28nm FD-SOI ASIC technology in order to scale both frequency and voltage simultaneously and accordingly, to achieve more power dissipation and energy reduction. The different blocks of the system were synthesized with Synopsys Design Compiler 2014.09 on a 28nm UTBB FD-SOI RVT standard cell library. The power estimations were performed with Synopsys PrimeTime 2013.06 assuming 20% of switching activity. In order to estimate the power consumption of the blocks at different voltages, the libraries have been characterized down to 0.5V with 0.1V steps using Cadence Liberate. The design was synthesized at 1.0V (slow-slow corner, 125C), while the signoff was performed at different operating voltages

Table 3
Power consumption and area utilization of the nodes synthesized on ASIC at the operating frequency of 500.0 MHz (0.9V, ss, 125C) and typical conditions (tt, 25C, 1V) for estimating the power numbers.

Node	Area [mm²]	Leakage Power @ 1V [mW]	Dynamic Power [mW] @ 1V, 500 MHz
N0	1.78	0.0298	51
N1	0.9	0.0146	24.75
N2	1.81	0.0292	48.75
N3	0.79	0.0136	10.91
N4	0.79	0.0130	15
N5	0.79	0.0128	19.65
N8	1.82	0.0294	48.6
Total	8.68	0.1424	218.66

Table 4
Impact of DVFS method on the dynamic power dissipation of the nodes as well as the entire platform on ASIC prototype at scaled down voltages and frequencies. DP stands for Dynamic Power.

| Node	DP [mW] @ 0.9V 480.0 MHz	DP [mW] @ 0.8V 366.0 MHz	DP [mW] @ 0.7V 238.0 MHz	DP [mW] @ 0.6V 119.0 MHz	DP [mW] @ 0.5V 55.0 MHz
N0	33.58	19.15	9.65	3.5	1.17
N1	16.3	9.29	4.68	1.7	0.57
N2	32.1	18.3	9.22	3.34	1.12
N3	7.18	4.09	2.06	0.75	0.25
N4	9.88	5.63	2.84	1.03	0.35
N5	12.94	7.38	3.72	1.35	0.45
N8	32	18.25	9.2	3.34	1.12
Total	149.96	85.51	43.09	15.62	5.23
Gain %	31.42	60.89	80.29	92.86	97.61

Table 5
Dynamic power dissipation of each node before applying FCS, after applying FCS with DFS and also with DVFS on ASIC.

Node	DP[mW] FCS Inactive, 200 MHz	DP [mW] FCS Active with DFS @ 1V	DP [mW] FCS Active with DVFS
N0	20.4	5.61	1.7
N1	9.9	9.9	9.9
N2	19.5	5.36	1.12
N3	4.36	1.2	0.25
N4	6	1.65	0.35
N5	7.86	2.16	0.45
N8	19.44	5.35	1.12
Total	87.46	31.23	14.89

Table 6
General comparison of the impact of DFS technique on FPGA and ASIC dynamic power dissipation at the same operating condition.

Node	Dynamic Power [mW] FPGA FCS Active with DFS	Dynamic Power [mW] ASIC FCS Active with DFS @ 1V
N0	284.43	5.61
N1	272.83	9.9
N2	391.23	5.36
N3	104.13	1.2
N4	114.61	1.65
N5	105.07	2.16
N8	344.61	5.35
Total	2092.6	31.23

(from 1.0V to 0.5V with steps of 0.1V), in order to provide to the software the knowledge of the maximum operating frequency and power consumption of the system at the different voltage supplies evaluated in this work. Table 3 shows the post-syntesis results in which the maximum achieved operating frequency is 500.0 MHz at nominal voltage in the slow corner (ss, 125 °C, 0.9V) and leakage power and dynamic power consumption are measured in typical operating conditions (tt, 25 °C, 1V). Similar to FPGA, the area occupation and also the leakage and dynamic power consumption of the CGRA nodes will be almost doubled by increasing the size of template-based CGRAs from CREMA (N1) to AVATAR (N0, N2 and N8).

Subsequent to synthesizing the whole platform on ASIC, we started to scale down both frequency and voltage simultaneously in a way that at each phase, the dynamic power consumption has been measured. However, the measurement of updated leakage power is ignored due to its negligible value. Node-by-node breakdown of dynamic power measurements at each stage with scaled-down frequencies and voltages is depicted in Table 4 where gains are calculated against dynamic power at 1V and 500.0 MHz. It can be observed that by scaling down the operating frequency and the voltage from 500.0 MHz and 1V to 480.0 MHz and 0.9V, respectively, the dynamic power dissipation is reduced by 31.42%. By moving forward in scaling down the operating frequency and the voltage down to 55 MHz and 0.5V, the dynamic power dissipation reduction showed 97.61% saving against to the first stage (500.0 MHz and 1V).

Then we kept the maximum operation frequency at the fixed value of 200.0 MHz, the same operating condition as FPGA, and
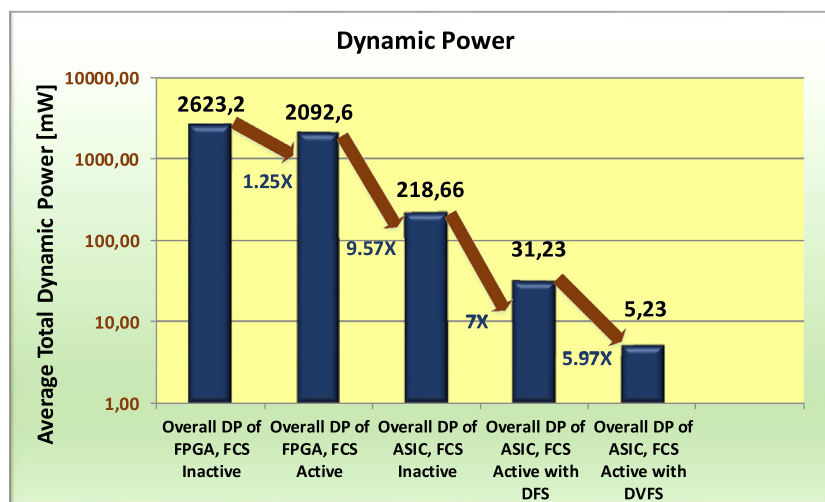
Fig. 8. General Comparison of Total Dynamic Power of FPGA Inactive FCS, FPGA Active FCS, ASIC Inactive FCS, ASIC Active FCS with DFS and with DVFS. DP stands for Dynamic Power.

started to estimate the dynamic power consumption in three following cases: inactive state of FCS, active state of FCS with DFS and also with DVFS. As it can be seen from Table 5, the dynamic power dissipation of N1, Frequency Offset Estimation block, as the worst-case candidate is constant in all three cases while the dynamic power dissipation of other nodes started to be reduced (in total by 64.29%) once the FCS starts to tune the operating frequency of the cores. Therefore, all the cores approached successfully to the equalization region, targeted automatically by FCS, while running at 55.0 MHz operating frequency. In order to get more benefits in terms of power mitigation, the supply voltages are also scaled down to 0.5V in parallel with the frequency scaling which resulted in mitigation of the total dynamic power dissipation by 82.98%.

5.3. Mixed comparison between FPGA and ASIC technologies

In order to have a fair comparison between the OFDM receiver implementation on two different technologies, FPGA and ASIC, the estimated dynamic power dissipation is compared in the case of 200.0 MHz maximum operating frequency for running the whole platform at the condition of active state of FCS with DFS. The results, depicted in Table 6, show the significant power mitigation of ASIC implementation. The general comparison as a visual summary of total dynamic power dissipation of FPGA and ASIC is also shown in Fig. 8 by considering the following cases: FPGA with inactive state of FCS, FPGA with active state of FCS, ASIC with inactive state of FCS, ASIC with active state of FCS and DFS and also ASIC with active state of FCS and DVFS. Applying FCS on FPGA resulted in dynamic power dissipation reduction by 1.25X while on ASIC the amount of achieved power saving with DFS is 7X which can be further reduced by 5.97X with DVFS. This approach paves the way for self-aware systems for energy efficient OFDM receiver, mapped on HARP, by mitigating the signal transition activity over the entire platform with reference to the worst-case performing core.

6. Conclusions

This paper presents the power mitigation of a Heterogeneous Accelerator-Rich Platform (HARP) on Stratix-V Field-Programmable Gate Array (FPGA) device and 28nm Ultra-Thin Body and Buried oxide (UTBB) Fully-Depleted Silicon-On-Insulator (FD-SOI) Application-Specific Integrated Circuit (ASIC) technology by employing a Dynamic Frequency and Voltage Scaling (DVFS) technique in an Orthogonal Frequency-Division Multiplexing (OFDM) receiver test case. The platform consists of three Reduced Instruction Set Computing (RISC) cores and four template-base Coarse-Grained Reconfigurable Arrays (CGRAs) nodes arranged over a Network-on-Chip (NoC). Template-based CGRAs are crafted with different sizes according to the algebraic expressions of computationally intensive tasks of OFDM receiver blocks. Due to the high-level of switching activity on the platform and high power dissipation of CGRAs, dynamically scaling down the operating frequency and the voltage for mitigating the power dissipation is vital. To face this issue, Feedback Control System (FCS) implemented in RISC software continuously monitors the performance of each CGRA and can be employed to tune the clock frequency of the CGRA nodes by selecting a worst-case execution time candidate. RISC processors are responsible to monitor the performance of their associated CGRA cores continuously and store the counted clock cycles in a reserved location of their data memory. FCS uses several iterations to degrade the performance of the CGRA nodes to the equalization region in an user-defined and application-specific range of frequencies. Furthermore, in some cases where the CGRAs are reconfigured at run-time for performing the new tasks, FCS can update itself automatically and define new performance equalization region in

order to make the balance between the computing-nodes. In the case of FPGA implementation, the dynamic power dissipation and total power dissipation of the platform showed 20.2% and 16.8% reduction, respectively, subsequent to applying the FCS technique with Dynamic Frequency Scaling (DFS). Furthermore, by moving to ASIC technology, both frequency and voltage have been scaled simultaneously which resulted in significant dynamic power reduction while the nodes approached the equalization region. Achieved results from prototyping the heterogeneous multicore architecture on FPGA and in particular ASIC proved that SElf-awarE Computing (SEEC) models such as FCS by the use of DFS/DVFS techniques can be practically and realistically fruitful as one step forward to mitigate the Dark Silicon issue by reducing the instantaneous power dissipation and therefore the heat dissipation.

References

[1] G.E. Moore, Cramming more components onto integrated circuits, Electronics (Basel) 38 (8) (1965).

[2] A. Pedram, S. Richardson, S. Galal, S. Kvatinsky, M. Horowitz, Dark memory and accelerator-rich system optimization in the dark silicon era, in: IEEE Design & Test, vol. 34, no. 2, 2017, pp. 39–50. doi: 10.1109/MDAT.2016.2573586.

[3] M.B. Taylor, J. Kim, J. Miller, D. Wentzlaff, F. Ghodrat, B. Greenwald, H. Hoffman, P. Johnson, L. Jae-Wook, W. Lee, A. Ma, A. Saraf, N. Seneski, N. Shnidman, V. Strumpen, M. Frank, S. Amarasinghe, A. Agarwal, The raw microprocessor: a computational fabric for software circuits and general-purpose programs, Micro, IEEE 22 (2) (2002). 25,35

[4] G. Venkatesh, J. Sampson, N. Goulding, S. Gracia, V. Bryksin, J.L. Martinez, S. Swanson, M.B. Taylor, Conservation cores: reducing the energy of mature computations, ASPLOS 10 (2010) 205–218.

[5] M.B. Taylor, Is dark silicon useful?: Harnessing the four horsemen of the coming dark silicon apocalypse, in: proceedings of the 49th Annual Design Automation Conference (DAC 12), ACM, NY, USA, pp. 1131–1136.

[6] A.K. Coskun, J.L. Ayala, D. Atienza, T.S. Rosing, Y. Leblebici, Dynamic Thermal Management in 3D Multicore Architectures, in: Proc. of Design, Automation & Test in Europe Conference & Exhibition, Nice, France, 2005, pp. 1410–1415.

[7] W. Kim, M. Gupta, G.Y. Weil, D. Brooks, System Level Anaysis of Fast, Per-core DVFS Using On-chip Switching Regulators, in: Proc. International Symposium on High Performance Computer Architecture (HPCA), 2008, pp. 123–134.

[8] H. Hoffmann, et al., Self-aware Computing in the Angstrom Processor, in: DAC Design Automation Conference 2012, San Francisco, CA, 2012, pp. 259–264. doi: 10.1145/2228360.2228409.

[9] W. Hussain, R. Airoldi, H. Hoffmann, T. Ahonen, J. Nurmi, HARP²: An X-scale Reconfigurable Accelerator-rich Platform for Massively-parallel Signal Processing Algorithms, Journal of Signal Processing Systems, Springer, 2015.

[10] W. Hussain, H. Hoffmann, T. Ahonen, J. Nurmi, Power Mitigation by Performance Equalization in a Heterogeneous Reconfigurable Multicore Architecture, in: Journal of Signal Processing Systems, Springer, 2017, pp. 287–297. doi: 10.1007/s11265-016-1142-5.

[11] D. Melpignano, L. Benini, E. Flamand, B. Jego, T. Lepley, G. Haugou, F. Clermidy, D. Dutoit, Platform 2012, a Many-core Computing Accelerator for Embedded Socs: Performance Evaluation of Visual Analytics Applications, in: Proc. 49th Annual Design Automation Conference (DAC '12), ACM, New York, NY, USA, 2012, pp. 1137–1142.

[12] F. Conti, C. Pilkington, A. Marongiu, L. Benini, He-p2012: Architectural Heterogeneity Exploration on a Scalable Many-core Platform, in: Proc. of the 24th edition of the great lakes symposium on VLSI (GLS- VLSI '14), ACM, New York, NY, USA, 2014, pp. 231–232.

[13] N.S. Voros, M. Hübner, J. Becker, M. Kühnle, F. Thomaitiv, A. Grasset, P. Brelet, P. Bonnot, F. Campi, E. Schüler, H. Sahlbach, S. Whitty, R. Ernst, E. Billich, C. Tischendorf, U. Heinkel, F. Ieromnimon, D. Kritharidis, A. Schneider, J. Knaeblein, W. Putzke-Röming, MORPHEUS: A heterogeneous dynamically reconfigurable platform for designing highly complex embedded systems, ACM Trans. Embed. Comput. Syst. 12 (3) (2013) 33.

[14] F. Thoma, M. Kuhnle, P. Bonnot, E.M. Panainte, K. Bertels, S. Goller, A. Schneider, S. Guyetant, E. Schuler, K.D. Muller-Glaser, J. Becker, MORPHEUS: Heterogeneous reconfigurable computing international conference on field programmable logic and applications, FPL (2007) (2007) 409–414. 27–29.

[15] D. Rossi, F. Campi, S. Spolzino, S. Pucillo, R. Guerrieri, A Heterogeneous Digital Signal Processor for Dynamically Reconfigurable Computing, in: IEEE Journal of Solid-State Circuits, vol. 45, no. 8, 2010, pp. 1615–1626. doi: 10.1109/JSSC.2010.2048149.

[16] F. Conti, et al., An iot endpoint system-on-chip for secure and energy-efficient near-sensor analytics, in: IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 64, no. 9, 2017, pp. 2481–2494. doi: 10.1109/TCSI.2017.2698019.

[17] H.J. Choi, Y.J. .Park, H.H. Lee, C.H. Kim, Adaptive dynamic frequency scaling for thermal-aware 3d multi-core processors, in: Computational Science and Its Applications-ICCSA 2012, Lecture Notes in Computer Science, 7336, 2012, pp. 602–612. ISBN: 978-3-642-31127-7.

[18] X. Chen, Z. Xu, H. Kim, P.V. Gratz, J. Hu, M. Kishinevsky, U. Ogras, R. Ayoub, Dynamic voltage and frequency scaling for shared resources in multicore processor designs, in: Proceedings of the 50th Annual Design Automation Conference (DAC 13). Article 114, ACM, NY, USA, 2013, p. 7. doi: 10.1145/2463209.2488874.

[19] S.M.A.H. Jafri, M.A. Tajammul, A. Hemani, K. Paul, J. Plosila, H. Tenhunen, Energy-aware-task-parallelism for efficient dynamic voltage, and frequency scaling, in: CGRAs, 2013 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS XIII), 2013, pp. 104–112. doi: 10.1109/SAMOS.2013.6621112.

[20] M.H. Haghbayan, A.M. Rahmani, A.Y. Weldezion, P. Liljeberg, J. Plosila, A. Jantsch, H. Tenhunen, Dark silicon aware power management for manycore systems under dynamic workloads, in: 2014 32nd IEEE International Conference on Computer Design (ICCD), 2014, pp. 509–512.

[21] Y. Sinangil, et al., A self-aware processor soc using energy monitors integrated into power converters for self-adaptation, in: Symposium on VLSI Circuits Digest of Technical Papers, Honolulu, HI, 2014, pp. 1–2. doi: 10.1109/VLSIC.2014.6858424.

[22] M. Möstl, J. Schlatow, R. Ernst, H. Hoffmann, A. Merchant, A. Shraer, Self-aware systems for the internet-of-things, in: 2016 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Pittsburgh, PA, 2016, pp. 1–9.

[23] R. Airoldi, F. Garzia, J. Nurmi, Improving reconfigurable hardware energy efficiency and robustness via DVFS-scaled homogeneous MP-soc, in: 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum, Shanghai, 2011, pp. 286–289. doi: 10.1109/IPDPS.2011.160.

[24] D. Rossi, et al., A self-aware architecture for PVT compensation and power nap in near threshold processors, in: IEEE Design & Test, vol. 34, no. 6, 2017, pp. 46–53. doi: 10.1109/MDAT.2017.2750907.

[25] A. Bartolini, R. Diversi, D. Cesarini, F. Beneventi, Self-aware thermal management for high performance computing processors, in: IEEE Design & Test, doi:10.1109/MDAT.2017.2774774.

[26] J. Kylliäinen, T. Ahonen, J. Nurmi, General-purpose embedded processor cores - the COFFEE RISC example, in: J. Nurmi (Ed.), Processor Design: System-on-Chip Computing for ASICs and FPGAs, Kluwer Academic Publishers / Springer Publishers, 2007, pp. 83–100. Ch. 5, ISBN-10: 1402055293, ISBN-13: 978-1-4020-5529-4.

[27] F. Garzia, W. Hussain, J. Nurmi, CREMA, a coarse-grain re-configurable array with mapping adaptiveness, in: Proc. 19th International Conference on Field Programmable Logic and Applications (FPL 2009), IEEE, Prague, Czech Republic, 2009.

[28] C. Brunelli, F. Garzia, C. Giliberto, J. Nurmi, A dedicated DMA logic addressing a time multiplexed memory to reduce the effects of the system buss bottleneck, in: Proc. 18th International Conference on Field Programmable Logic and Applications, (FPL 2008), Heidelberg, Germany, 2008, pp. 487–490.

[29] S. Nouri, W. Hussain, J. Nurmi, Evaluation of a heterogeneous multicore architecture by design and test of an OFDM receiver, in: IEEE Transactions on Parallel and Distributed Systems, 2017, p. 1722. doi: 10.1109/TPDS.2017.2706691.

[30] S. Nouri, J. Nurmi, Power mitigation of a heterogeneous multicore architecture by frequency scaling in an OFDM receiver test case, in: Nordic Circuits and Systems Conference (NORCAS), Linkoping, Sweden, 2017, pp. 23–25. doi: 10.1109/NORCHIP.2017.8124987.

Sajjad Nouri is working as a Researcher and studying as a Doctor of Technology student in Laboratory of Electronics and Communications Engineering, Tampere University of Technology (TUT), Tampere, Finland. He has a B.Sc degree in Software Engineering from the University of Guilan, Iran. He received his M.Sc degree with distinction in Information Technology in June 2015 from TUT. His research work contains design and implementation of accelerators specialized for Software Defined Radio (SDR) applications by using different template-based Coarse-Grained Reconfigurable Arrays (CGRAs) and mapping their VHDL model onto the FPGA for evaluating the designed accelerators in terms of different performance metrics. He is also working on Accelerator-Rich Architectures and Heterogeneous Multicore Platforms in order to maximize the number of computational resources for exploiting the unutilized part of the chip which is known as Dark Silicon. He also has been a Visiting Scientist at Ruhr-University Bochum (RUB), Bochum, Germany.

Davide Rossi, received the PhD from the University of Bologna, Italy, in 2012. He has been a post doc researcher in the Department of Electrical, Electronic and Information Engineering Guglielmo Marconi at the University of Bologna since 2015, where he currently holds an assistant professor position. His research interests focus on energy efficient digital architectures in the domain of heterogeneous and reconfigurable multi and many-core systems on a chip. This includes architectures, design implementation strategies, and runtime support to address performance, energy efficiency, and reliability issues of both high end embedded platforms and ultra-low-power computing platforms targeting the IoT domain. In these fields he has published more than 60 paper in international peer-reviewed conferences and journals.

D.Sc.(Tech) Jari Nurmi works as a Professor at Tampere University of Technology, Finland since 1999, in the Faculty of Computing and Electrical Engineering. He is working on embedded computing systems, System-on-Chip, wireless localization, positioning receiver prototyping, and software-defined radio. He held various research, education and management positions at TUT since 1987 (e.g. Acting Associate Professor 1991–1994) and was the Vice President of the SME VLSI Solution Oy 1995–1998. Since 2013 he is also a partner and co-founder of Ekin Labs Oy, a research spin-off company, now headquartered in Silicon Valley as Radiomaze, Inc. He has supervised 19 PhD and over 130 MSc theses at TUT, and been the opponent or reviewer of 33 PhD theses for other universities worldwide. He is a senior member of IEEE, and member of the technical committee on VLSI Systems and Applications at IEEE CAS. In 2004, he was one of the recipients of Nokia Educational Award, and the recipient of Tampere Congress Award in 2005. In 2011 he received IIDA Innovation Award, and in 2013 the Scientific Congress Award and HiPEAC Technology Transfer Award. He is a steering committee member of four international conferences (chairman in two). He has edited 5 Springer books, and has published over 350 international conference and journal articles and book chapters.