The First 25 Years of the FPL Conference – Significant Papers

Philip H.W. Leong, The University of Sydney
Hideharu Amano, Keio University
Jason Anderson, University of Toronto
Koen Bertels, Delft University of Technology
João M.P. Cardoso, Universidade do Porto
Oliver Diessel, UNSW Australia
Guy Gogniat, University of South Brittany
Mike Hutton, Altera Corp
JunKyu Lee, The University of Sydney
Wayne Luk, Imperial College London
Patrick Lysaght, Xilinx Research Labs
Marco Platzner, University of Paderborn
Viktor K. Prasanna, University of Southern California
Tero Rissan, Nokia Corporation
Cristina Silvano, Politecnico di Milano
Hayden Kwok-Hay So, University of Hong Kong
Yu Wang, Tsinghua University

CCS Concepts: *Computer systems organization → Reconfigurable computing; Hardware → Reconfigurable logic and FPGAs;

1. INTRODUCTION

The first International Conference on Field-Programmable Logic and Applications (FPL) was held in 1991 at Oxford University. In the ensuing years, it has become the largest meeting on field-programmable gate array (FPGA) technologies and systems, and many important contributions have been published at the conference.

This work was supported under the Australian Research Councils Linkage Projects funding scheme (project number LP130101034), and the Faculty of Engineering and Information Technologies, The University of Sydney by the Faculty Research Cluster Program.

Authors' addresses: Philip Leong and JunKyu Lee, The University of Sydney, Australia, {philip.leong,jun.kyu.lee}@sydney.edu.au; Hideharu Amano, Keio University, Japan, hunga@am.ics.keio.ac.jp; Jason Anderson, University of Toronto, Canada, janders@ece.utoronto.ca; Koen Bertels, Delft University of Technology, Netherlands, K.L.M.Bertels@tudelft.nl; João M.P. Cardoso, Universidade do Porto, Portugal, jmpc@acm.org; Oliver Diessel, University of NSW, Australia, odiesel@csse.unsw.edu.au; Guy Gogniat, University of South Brittany, France, guy.gogniat@univ-ubs.fr; Mike Hutton, Altera Corp, San Jose, USA, mhutton@altera.com; Wayne Luk, Imperial College London, UK, w8/doc.ic.ac.uk; Patrick Lysaght, Xilinx Research Labs, San Jose, USA, patrick.lysaght@xilinx.com; Marco Platzner, University of Paderborn, Germany, platzner@upb.de; Viktor Prasanna, University of Southern California, USA, prasanna@usc.edu; Tero Rissan, Nokia, Finland, tero.riass@nokia.com; Cristina Silvano, Politecnico di Milano, Italy, cristina.silvano@polimi.it; Hayden Kwok-Hay So, University of Hong Kong, Hong Kong, hso@eee.hku.hk; Yu Wang, Tsinghua University, China, yu-wang@tsinghua.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2016 ACM. 1936-7406/2016/12-ART123456 $15.00
DOI: XXXXXXX.XYYYYYY

ACM Transactions on Reconfigurable Technology and Systems, Vol. 0, No. 0, Article 123456, Publication date: December 2016.
This paper lists the most significant contributions from 1991–2014 [Leong et al. 2015]. The selection was made by an international Significant Papers Committee (SPC), comprising the authors of this paper. Only regular papers were considered and Google Scholar was used for citation counts. Since an objective process was desired, those papers having more than 9 citations per year, or more than 100 overall citations were included on an initial shortlist. Concurrently an open call to the community was made through the FPGA mailing list (fpga-list@mailman.sydney.edu.au), inviting nomination of papers with impact other than the number of citations. A total of 24 nominations were received and all papers that were not already on the initial shortlist (13 papers) were included in the final shortlist.

The resulting 60 shortlisted papers were divided into 5 categories, each with a corresponding subcommittee. SPC members served on a subcommittee according to their expertise. Since some nominated papers were authored by SPC members, subcommittee Chairs were chosen to have no conflict of interest with any of the papers within, and tasked with ensuring that no SPC member had conflicts of interest during the paper selection process. Following discussions within each subcommittee, recommendations were consolidated for the SPC. A final dialogue regarding the choices was then undertaken.

The total number of papers considered was 1,765 (this figure may be larger than the total number of full papers as it was not possible to distinguish them from posters in some cases). The process described above resulted in 27 papers (1.5% of the total) being selected. Inevitably, the inclusion and exclusion of certain papers will be the subject of debate; however, this list represents the best efforts of the SPC to ensure a fair and objective selection process.

2. FPL SIGNIFICANT PAPER LIST (1991-2014)

2.1. Applications and Benchmarks

Software executing on instruction-driven processors such as multicores or graphics processing units (GPUs) face limitations on computational speed and energy efficiency due to the need for extensive logic to fetch and decode instructions. Data-driven FPGA designs address these issues through higher degrees of customisation, allowing more operations to be performed per clock cycle enabling improved speed, energy efficiency and ability to respond to real-time events.

Eight applications papers are introduced, each of which has influenced the development of other FPGA-based applications and benchmarks. FPGAs are well suited to signal processing applications in which multiplication often involves a constant multiplicand, and customisation in this case can save hardware resources [Petersen and Hutchings 1995], and real-time operation, particular in video-signal processing has always been a challenge [Haynes et al. 1999]. DNA sequence alignment is another computationally intensive kernel, and customisation to allow more computational elements to fit on a single FPGA was described [Yu et al. 2003]. Another application which takes customisation to the extreme is to exploit process variations to produce physically uncloneable functions [Guajardo et al. 2007]. Arithmetic optimisation utilising arithmetic of different precisions has been effectively used to solve global atmospheric equations [Gan et al. 2013]. The general problem of multitasking on FPGAs has been extensively researched and introducing the restriction that only one task can operate at any time has important advantages [Simmler et al. 2000]. Two other benchmarking papers were included. A quantitative comparison between FPGAs, GPUs and CPUs for image processing applications [Asano et al. 2009], and a comparison of the round
2 SHA–3 candidates which played an important role in the competition of the standard [Baldwin et al. 2010].

This early work compared serial, parallel and distributed FPGA multiplier performance with that of digital signal processor (DSP) and application specific integrated circuit approaches. Their performance in finite impulse response (FIR) filters and a radix-4 fast Fourier transform was studied. The paper showed the potential of FPGAs in signal processing and laid the foundations for this key application domain. The authors showed how arithmetic units could be assembled on such devices, which were very rudimentary at that time. The key optimisation of constant operand multiplication was applied, and the potential for accelerating DSP problems on an emerging hardware platform identified. Many existing applications apply ideas from this work [Meyer-Baese and Meyer-Baese 2007], and research in this domain is ongoing.

– SONIC - a plug-in architecture for video processing [Haynes et al. 1999].
This work proposed a reconfigurable computing architecture for video processing which was comprised of multiple Plug-In Processing Elements (PIPEs) consisting of a Pipe Engine, Pipe Router and Pipe Memory. An application programming interface to handle resource allocation and scheduling was also introduced. Performance was demonstrated with a 19-tap 2D FIR filter with 8 PIPEs. This achieved more than 15 frames per second operating on $512 \times 512$ resolution video transferred over the PCI bus. This paper pioneered a reconfigurable hardware architecture with software plug-ins for accelerating video image processing applications and utilised modularity to reduce application design time. It had significant impact on reconfigurable media processing applications, hardware/software code design researches, and embedded video image processing research [Galuzzi and Bertels 2011]. It provides the foundation for the UltraSonic system [Haynes et al. 2002], which was commercialised by Sony Broadcast.

– Multitasking on FPGA Coprocessors [Simmler et al. 2000].
This paper details the requirements and overheads associated with preemptive multitasking for FPGA coprocessors. The proposed technique involved a single hardware task being executed at any time, this having three advantages: (1) while parallel execution of several tasks on the FPGA is possible, executing a single task at a time allows the full communication data rate with the host processor; (2) multitasking allows switching of the FPGA at any time and (3) I/O resources do not need to be shared. A task manager was also described, this being used to manage loading of circuits and states for different hardware tasks. This paper is significant in that it pioneers preemptive multitasking with an FPGA coprocessor and also inspired other multitasking research [Steiger et al. 2004a].

– A Smith-Waterman Systolic Cell [Yu et al. 2003].
This paper presents an improved systolic processing element (PE) for implementing the Smith-Waterman DNA sequence alignment algorithm on an FPGA. The authors proposed a scheme in which two PEs were merged into a compact cell, a technique commonly used in VLSI layout. Only by considering two PEs together could adjacent memory elements be utilised, this serving to achieve a 25% area savings. A large linear systolic array of elements was connected together to achieve a high degree of parallelism, and it was interfaced to an FPGA board via the SDRAM bus to achieve the highest reported density and system performance at the time.
Sequence alignment is the computational bottleneck in many bioinformatics applications including sequence database searching, multiple sequence alignment, genome assembly and short read mapping. Thus acceleration of bioinformatics applications using FPGAs has become an important application domain and the Smith-Waterman algorithm continues to be used. Compared with previous designs, this work removed the requirement of runtime reconfiguration while maintaining equivalent functionality. The paper represents a reference design widely used in that domain [Gokhale and Graham 2006] and inspired other work in FPGA-based pattern matching [Van Court and Herbordt 2007] and biosequence analysis [Oliver et al. 2005].

– Physical Unclonable Functions, FPGAs and Public-Key Crypto for IP Protection [Guajardo et al. 2007].
This paper presented several new protocols to address the IP protection problem on FPGAs, supporting hardware authentication, hardware platform authentication and complete design confidentiality. The authors’ approach relies on public-key cryptography, with secret key information being generated by a physical uncloneable function (PUF) implemented in the FPGA. Since a unique PUF output is generated internal to the device, security is enhanced by removing the requirement to transfer the key on or off the FPGA. The paper also details how the PUF can be implemented using the random initialization state of on-FPGA static RAM cells.
This paper provides efficient protocols for IP protection and introduces an original solution for PUFs based on SRAM memories with good properties. It has become one of the key references in this area [Maes and Verbauwhede 2010].

– Performance comparison of FPGA, GPU and CPU in image processing [Asano et al. 2009].
This paper gives a quantitative performance comparisons between FPGA, GPU, and CPU implementations of image processing applications. Special characteristics for each platform were highlighted, e.g. an FPGA can implement an optimized circuit according to an application, a GPU can utilize many cores, and a CPU can utilize its SIMD features. It found that in two-dimensional filters, GPUs are the best choice, since shared variables are not required. In stereo-vision and K-means clustering applications, an FPGA excels because it avoids local memory access conflicts. Finally, a CPU is usually a good choice for applications in which cache miss rates are low.
This research provided informative descriptions of how performance is affected by application characteristics. It has significance from a tutorial perspective and impacted other performance comparison research [Pauwels et al. 2012].

– FPGA Implementations of the Round Two SHA-3 Candidates [Baldwin et al. 2010].
This paper provides a summary of the algorithms associated with the round 2 candidates for the secure hash algorithm 3 (SHA-3), along with a detailed study of their suitability for an efficient hardware implementation. Hardware implementations of all of the algorithms were compared for different message sizes. Moreover, this was the only work to include padding as part of the hardware. At the time of publication, the US National Institute of Standards and Technology (NIST) was holding a competition to determine a successor to the SHA-2 algorithm.
The research work outlined in this paper was considered as part of the SHA-3 algorithm selection process by NIST [Turan et al. 2011]. The Keccak candidate was identified in the paper as having the highest overall performance and best performance per area. Keccak was subsequently chosen for the SHA-3 hashing standard in August 2015 [NIST 2015].

This paper showed that reconfigurable supercomputing can provide faster and more energy efficient solutions, compared to conventional supercomputers for global climate modeling. The paper also provides guidance on how to apply mixed-precision arithmetic to complex problems that do not easily fit onto limited hardware resources. The experimental results reported two orders of magnitude speedup over a multicore CPU, and the design was one order of magnitude faster and one order of magnitude lower power than a CPU-GPU hybrid.

It has been cited by a number of promising research efforts that follow a similar direction, including [Düben and Palmer 2014] and [Russell et al. 2015].

2.2. Architecture

The development of reconfigurable architectures to efficiently perform computing tasks has been studied since the introduction of FPGAs. Common issues including striking a balance between fine and coarse-grained computing resources, time-multiplexing to efficiently utilize resources, interfacing with processors, and reducing power and energy consumption.

Five architecture papers are introduced in this subsection. The RaPiD architecture combined static and cycle-by-cycle dynamic configuration [Ebeling et al. 1996]. SCORE introduced the abstraction of streams of computation, mapped to a programmable fabric via time-multiplexing [Caspi et al. 2000]. The ADRES approach combined a very-long instruction word processor with coarse-grained arithmetic blocks [Mei et al. 2003]. One of the advantages of FPGAs is in power and energy consumption. Power modelling for FPGAs has been a major concern and an early paper largely solved this problem in a general manner for the widely used VPR tool [Poon et al. 2002]. Finally, using multiple VDD supplies allows power to be reduced without sacrificing performance [Gayasen et al. 2004].

- RaPiD - reconfigurable pipelined datapath [Ebeling et al. 1996].

This early paper proposed a reconfigurable, deeply pipelined datapath tailored to application specific requirements. Pipelined datapaths consisting of a mixture of arithmetic logic units (ALUs), multipliers, general-purpose registers and local memories are constructed by a combination of static and dynamic configurations. While the static configuration determines the overall pipeline structure through a set of programmable routing switches, the dynamic configuration determines the cycle-by-cycle operations of the pipeline similar to the operation of a microcontroller.

This paper was highly influential, having inspired numerous subsequent works in pipelined and coarse-grained reconfigurable architectures [Baumgarte et al. 2003]. Its pipelined reconfiguration has influenced commercial products such as Tabula, and many ideas originally developed for the RaPiD design have also found applications in application specific standard products and other tightly-coupled accelerator architectures designs. As this work dates back to the 90s, it was truly visionary and fundamental to the evolution of field-programmable architectures.

- Stream Computations Organized for Reconfigurable Execution (SCORE) [Caspi et al. 2000].

This paper introduces a new computational model for reconfigurable computing, based upon streams of computation virtualized onto a programmable fabric via time-multiplexing. The model facilitates a language abstraction and mapping that allows arbitrary computations to be implemented onto a target device without recompilation, overcoming the key challenge of vendor and device generation portability, and the con-
constraint of fixed resources on the physical hardware. The streaming/paging abstraction introduced in the paper offers solutions for important practical problems such as dealing with exceptions via flow-control. SCORE continued to grow and develop beyond this extended abstract, and a summary can be found in [DeHon et al. 2006].

The formalized model in SCORE influenced subsequent work on overlays, hardware virtualization and heterogeneous computing. Moreover, virtualization has become one of the largest growth areas for FPGAs, which are now moving into datacenters and cloud computing as compute accelerators for processors. This paper provided mechanisms to address many of the key problems: bitstream portability, targeting heterogeneous resources, scaling computational resources to match availability on the physical fabric, and tolerating uncertain latency on links between operators. It not only addressed interesting problems of the day but foreshadowed those of the future. The language and model proved valuable for other resource- and performance- scaleable implementations, including custom VLIW mappings [Kapre and DeHon 2011].

The streaming model plays a prominent role in some of the more usable commercial software offerings for reconfigurable computers including Ambric’s Structural Object Programming Model [Butts et al. 2007] and IBM’s Lime [Auerbach et al. 2010].

– A Flexible Power Model for FPGAs [Poon et al. 2002].

This was the first work to incorporate a power model into the widely used open-source VPR FPGA architecture and CAD framework. While today’s FPGAs incorporate power-aware features, and news abounds about FPGAs in low-power/mobile scenarios, this power-modeling work was done at a time when power was not a significant axis along which FPGAs were optimized. The proposed power model is flexible in the sense that, instead of modeling the power of a specific FPGA, the model is adaptive to the architecture specified to VPR. Dynamic and static power can be estimated for a range of LUT sizes, cluster sizes, and interconnection structures.

The “Poon power model” has been used and improved [Li et al. 2005] by researchers in a variety of power-aware CAD and architecture research endeavours. The power model, its associated source code, and documentation were released into the public domain, and for several years it was the primary model used to study power consequences of architecture and CAD algorithm changes. This work thus represents a significant service to the FPGA community.


This paper introduces a novel coarse-grained reconfigurable array (CGRA), which is a style of FPGA that uses large ALU-like logic blocks and multi-bit routing. While the CGRA concept was already known at the time of publication, the ADRES (Architecture for Dynamically Reconfigurable Embedded System) architecture incorporates features not seen in prior work. The CGRA fabric in ADRES is closely coupled with a Very-Long Instruction Word (VLIW) processor, and in fact is able to access the same register file as the processor, obviating the need for costly transfers to/from the CGRA. VLIW processors frequently use a concept called loop pipelining to exploit loop-level parallelism by overlapping adjacent loop iterations. With the attached CGRA fabric, ADRES offers the added ability to exploit spatial parallelism, providing higher computational throughput than the VLIW processor alone. The functional units in the VLIW processor are also part of the CGRA fabric, permitting resources to be shared between the two.

The ADRES architecture has had a far-reaching impact in the commercial space, which took many years to materialize. Samsung Electronics began a long-term collaboration with the ADRES researchers to develop a commercial CGRA, which today is known...
as the Samsung Reconfigurable Processor (SRP). SRP publications began to appear in the 2010s; it has been used in a variety of commercial Samsung products, including smartphones. ADRES can be viewed as an ancestor of the SRP [Kim et al. 2012].

– **A Dual-VDD Low Power FPGA Architecture** [Gayasen et al. 2004]. This work proposed a CAD-driven approach to configure the power domains of different portions of the FPGA. Both logic and interconnect were optimised [Anderson and Najm 2006], by utilising slack from routed nets to assign voltage configurations. Similar approaches have been adopted commercially for power reduction using configurable back bias interacting with CAD tools in Altera Programmable Power Technology [Altera 2007] and by reconfigurable circuit fabrics from Xilinx [Tuan et al. 2012]. This work also helped inspire several academic efforts that considered interactions between configurable circuit fabrics and CAD flow optimizations for power reduction.

### 2.3. Design Methods and Tools

FPGA users often need to master complex design flows and evaluate a large design space to meet requirements. Advances in design methods and tools are thus of paramount importance for both researchers and practitioners, this continuing to be a crucial research area.

The four papers selected point to the breadth of research presented at the conference. They focus on placement and routing tools and algorithms [Betz and Rose 1997], generation of stream-based application-specific architectures from object-oriented descriptions [Mencer et al. 2000], an approach proposing power/energy reductions via the insertion of pipelining stages [Wilton et al. 2004], and FPGA support and a methodology for making easier the development of designs taking advantage of dynamic reconfiguration features [Lysaght et al. 2006].

– **VPR: A new packing, placement and routing tool for FPGA research** [Betz and Rose 1997]. This paper presented the Versatile Place and Route (VPR) tool which performs placement and routing. A key feature is that it targeted a parameterised island-style FPGA defined through a configuration file, allowing it to be adapted to a broad range of different architectures. Its simulated annealing based place and route algorithm had smaller routing area than all previous work and the paper introduced a new set of benchmark circuits.

VPR became an important piece of infrastructure for FPGA architecture and CAD tool research and has been widely used and extended by other research groups. This tool was further developed by the authors at Right Track CAD Corp which was acquired by Altera in 2000, and VPR then became the basis for both Altera's placement and routing and architecture evaluation tools.

This paper made a seminal contribution with the idea of a succinct and human-readable FPGA architecture description that drives a flexible CAD flow in order to explore new FPGA architectures, and has been used in a very wide range of FPGA architecture studies. The VPR architecture description language and CAD flow have been extended and enhanced in many ways, most recently with version 7.0 of the Verilog-to-Routing CAD flow [Luu et al. 2014]. VTR 7.0 can investigate the architecture not only of programmable logic and routing, as this original VPR could, but also arithmetic structures, RAM and DSP blocks and other special purpose structures. An enhanced version of VPR also forms the basis of Altera's FPGA Modeling Toolkit (FMT), which has been used to architect all of Altera's FPGAs from 2000 to the present day [Lewis et al. 2013]. VPR has also been used as a framework for the investigation of enhanced placement and routing algorithms both in academia and in Altera's commercial Quar-
tus II CAD suite which uses an enhanced version of VPR as its placement and routing engine [Ludwin and Betz 2011].

– **StReAm: Object-Oriented Programming of Stream Architectures Using PAM-Blox** [Mencer et al. 2000].
This paper presents StReAm, a domain specific compiler focused on compiling streaming-oriented applications to FPGA-based architectures. The approach was based on the PAM-Blox object-oriented module generation environment and employed a pipelined dataflow graph mapped directly to hardware. The advantage of this approach is that distributed FIFO buffers are able to exploit temporal data locality, while streaming data through pipelines exploits spatial data locality. StReAm uses operator overloading and template functions in C++ to create dataflow graphs which are consecutively scheduled to obtain a stream architecture.

They key contribution of this paper lies in an efficient compilation of a streaming-oriented programming model to FPGAs, this model being evident in contemporary products by Maxeler, a company founded by the authors. Maxeler has become the major provider of FPGA-based solutions in finance and oil related applications [Lindtjorn et al. 2011].

– **The Impact of Pipelining on Energy per Operation in Field-Programmable Gate Arrays** [Wilton et al. 2004].
This paper was the first to quantitatively study the impact of pipelining stages on power and energy consumption. It was found that pipelining could reduce energy per operation by between 40% and 90% on an Altera high-end FPGA, and a Xilinx low-cost FPGA. In addition, the authors showed that power-aware clustering enables further reductions, and interaction between pipelining and clustering was also studied to understand how the degree of pipelining affects the effectiveness of the lower-level CAD algorithms in reducing energy.

The quantitative results presented for two representative FPGA devices include real measurements and estimations provided by vendor-specific tools. The results presented had many implications and highlighted the importance to consider pipelining stages when datapath design and/or generation are addressed in contemporary FPGAs. This work foresaw the importance of pipelining for reducing energy, a fact which is employed in the recently announced Altera Stratix 10 [Hutton 2015].

– **Enhanced Architectures, Design Methodologies and CAD Tools for Dynamic Reconfiguration of Xilinx FPGAs** [Lysaght et al. 2006].
This paper presents Xilinx FPGA enhancements for providing better support for the creation of dynamically reconfigurable designs. Architecturally, it advocated pre-routed bus macros for implementing the communication ports to interface static and dynamic regions of a design, described new features of the Virtex 4 architecture to support finer reconfiguration granularity, and introduced a higher bandwidth reconfiguration port. It also presented a design flow which supported partially reconfigurable designs, enabled by the two key enhancements: allowing arbitrary rectangular regions to be reconfigured, and permitting signals in the static design to cross through partially reconfigurable regions without the use of a bus macro.

This paper was the first from a vendor to provide a complete dynamically reconfigurable FPGA and design flow. All major vendors now support similar flows [Kohn 2015; Altera 2010].
2.4. Dynamic Reconfiguration

Dynamic reconfiguration is the modification, at runtime, of a section of an FPGAs configuration memory with the aim of changing the implemented functionality. Reasons for doing so include: altered functional requirements; reducing energy consumption; reducing device size; correcting faults; and enhancing performance. As evidenced by the number of citations and publication dates, FPL is one of the most important venues for discussing the opportunities and challenges presented by dynamically reconfigurable FPGA-based systems. Six papers on dynamic reconfiguration were selected.

The six selected papers cover the early exploration of dynamically reconfigurable systems [Lysaght and Dunlop 1994], first efforts to model infrastructure requirements in a general manner [Brebner 1996], hardware virtualisation on a coarse-grained reconfigurable array and corresponding co-simulation techniques [Enzler et al. 2003], techniques to save and restore context to facilitate task pre-emption and re-location [Kalte and Porrmann 2005], methods and tools for generating dynamically reconfigurable systems [Koch et al. 2008], and applications to simultaneously error scrub and reconfigure the functionality of a design [Heiner et al. 2009].

– Dynamic reconfiguration of FPGAs [Lysaght and Dunlop 1994].
This paper comprehensively summarized the state of the art and the ideas underpinning dynamically reconfigurable systems at a very early stage in their history. The concepts that were explained set the scene for all subsequent investigation. The paper describes a large number of ideas that have later (in some cases much later) been the topic of significant papers themselves. It is a highly educational paper rather than being purely technical, yet it still includes a system example and discusses implementation issues. The limitations of the technology are also discussed.

This influential paper presented the basis for dynamic reconfiguration of FPGAs [Kohn 2015; Altera 2010]. It includes various concepts that are in use today, including a nomenclature for reconfiguration, FPGA-centric self-adapting systems [Sander et al. 2008], and the concept of logic caching. Many researchers have worked on solving or reducing the limitations identified in this paper, particularly optimizing FPGAs for speed of reconfiguration, state readback and update, automatic partitioning based on temporal specifications and tools for modeling new FPGA architectures.

– A virtual hardware operating system for the Xilinx XC6200 [Brebner 1996].
This paper introduced the concept of sea of accelerators and parallel harness virtual hardware models to support time-multiplexed hardware in which designs larger than physical FPGA resources could be executed. Swappable logic units (SLUs) were introduced, these being variable-sized processing elements that could be implemented on the FPGA resources and swapped out when needed. At the time of this paper, the Xilinx XC6200 was a new device which supported fine-grained partial reconfiguration. The sea of accelerators model was demonstrated along with an operating system which managed the configuration and utilisation of SLUs.

Concepts and challenges in developing an operating system for a reconfigurable computer were introduced and in the ensuing years, many further developments were inspired by working to address the challenges identified in this paper [Steiger et al. 2004b].

– Virtualizing Hardware with Multi-context Reconfigurable Arrays [Enzler et al. 2003].
The main contribution of this paper is an approach to overcome hardware resource limitations via hardware virtualization, enabling forward compatibility for reconfigurable fabrics. To achieve these goals, the paper proposes a coarse-grained, multi-context re-
configurable array that is attached as a co-processor to a CPU. The coarse-grained array is tailored to processing data streams with a macro-pipelined data path. If sufficient resources are available, pipeline stages are mapped spatially. Otherwise, time-multiplexing on a single array instance is used to provide virtualization allowing to execute macro-pipelines of arbitrary length on restricted hardware resources. The second contribution of this work was that it was one of the first works that proposed cycle-accurate co-simulation of the complete reconfigurable computing system to account for configuration, data-transfer and execution times. This detailed level of simulation allows the designer to accurately assess the performance of a single instance of the architecture and to perform a system-level evaluation for comparing design alternatives.

The impact and commercial relevance of this work is demonstrated by the considerable number of times it has been cited as prior art in patents, e.g. [Rohe et al. 2007].

– Context Saving and Restoring for Multitasking in Reconfigurable Systems [Kalte and Porrmann 2005].

This paper described a method for reducing dynamic task deallocation/relocation time by only saving the context of the task that is to be (re)moved. The idea was demonstrated using Xilinx Virtex FPGAs which supports column-wise reconfiguration. All modules were full height with varying width, reducing the placement problem to 1 dimension. Communication infrastructure is horizontal, with static modules in their own reserved area. Context relocation involved a configuration manager, a state extraction filter, state inclusion filter and REPLICA filter for relocation, with a database being used to store hardware task information. A detailed quantitative study of relocation time is made in the paper.

The idea and technique presented in the paper is at the core of any runtime system for dynamically reconfigurable computing. The solution was practical and formed the basis for future work in multitasking hardware operating systems [Lübbers and Platzner 2009].

– ReCoBus-Builder - A novel tool and technique to build statically and dynamically reconfigurable systems for FPGAs [Koch et al. 2008].

The paper presents an easily usable tool chain for implementing dynamically reconfigurable systems on FPGAs. ReCoBus-Builder provides sophisticated communication architectures allowing for direct bus connection as well as point-to-point connections. With this tool, over a hundred partially reconfigurable modules can be placed into a shared reconfigurable region. Here, modules can be relocated and instantiated multiple times while still being able to carry out dedicated communication with each module. This provides outstanding flexibility and is essential for reducing fragmentation in reconfigurable systems. The ReCoBus-Builder approach is further applicable for a component-based design methodology where modules from a library can be plugged together by a process called bitstream linking which can fully eliminate long synthesis runs or place and route steps.

ReCoBus-Builder was used for implementing several very sophisticated reconfigurable systems. For instance, in [Oetken et al. 2010], a system is presented where various 2-dimensionally relocatable reconfigurable video processing modules directly access external DDR memory using fast 64-bit DMA transfer concurrently with a complex PowerPC SoC on a Virtex-II Pro FPGA. The development of ReCoBus-Builder continued with the GoAhead tool [Beckhoff et al. 2012], which supports recent Xilinx FPGAs and which provides a powerful scripting interface. The currently developed next release of GoAhead will provide a TCL interface to the Vivado suite from Xilinx.
Many critical FPGA-based systems rely on configuration memory scrubbing to repair radiation-induced errors and a variety of configuration scrubbers have been proposed. One disadvantage of performing continuous configuration scrubbing is it occupies the configuration port, preventing the ability to perform other configuration functions such as partial reconfiguration. This paper proposed a simple method for performing partial reconfiguration in the course of a typical scrub cycle, thereby eliminating the need to interrupt scrubbing while the configuration port is used to perform a partial reconfiguration. To achieve this, the paper introduced a self scrubber, utilizing a small portion of the FPGA which performs the necessary operations to reconfigure a portion of the design while continuously scrubbing the entire FPGA. While the proposed technique is simple in nature, the paper’s significance arises from the interesting and clearly presented concepts, the convincing experimental evaluation, and the fact that the ideas have high practical value. Radiation-induced errors are gaining increasing attention due to their increased presence in modern FPGAs.

A number of interesting projects have been developed based on some of the ideas presented in this paper. A generic light-weight configuration controller was developed using ideas in this paper to provide local fault detection and isolation [Straka et al. 2013]. A novel configuration controller was created to facilitate FPGA bootstrapping for a multi-FPGA configuration process [Ostler et al. 2011]. A self-adaptive FPGA system was developed to modify the implementation of a function in run-time using both partial reconfiguration and configuration scrubbing to address changes in a radiation environment [Glein et al. 2014].

2.5. Security and Network-on-Chip

Networking has always been an important application domain since FPGAs are the only programmable technology which can keep up with network line rates, and they can combine on-chip networking hardware with custom pattern matching. Furthermore, the success of off-chip networking has raised the question of whether FPGAs should contain soft or hard networks on-chip (NoC). Advantages are that this can reduce required FPGA routing resources, introduce greater routing flexibility, and facilitate dynamic reconfiguration.

Four significant papers were selected around this topic. Two early papers identified the advantages of network intrusion detection on FPGAs and described designs which achieved performance far exceeding processor-based implementations [Gokhale et al. 2002; Sourdis and Pnevmatikatos 2003]. Significant papers concerning NoCs include one which advocates using them to support reconfigurable operating systems [Marescaux et al. 2003], and another quantifying the advantages of hardware NoCs for performance and energy efficiency [Abdelfattah and Betz 2013].

– Granidt: Towards Gigabit Rate Network Intrusion Detection Technology [Gokhale et al. 2002].

This paper described an FPGA-accelerated network intrusion detection system (NIDS). A compiler which translated filter rules from the public-domain Snort database into content addressable memories (CAMs) implemented on FPGA hardware, and a software representation of the rules used as an input to the software. The hardware appends the result of the match to the packet, and the rule software performs the action associated with the matching rules. The approach allowed changes to the rules to be implemented by creating new CAM entries, avoiding the need for reconfiguration. Using this approach, it was shown that the design could process data at a line rate of 2 Gbps, this being 24.9x the speed of the Snort software implementation.
This paper was seminal in FPGA-based network filtering technology, and has been cited by other influential papers such as [Kumar et al. 2006], and [Clark and Schimmel 2004b]. It heralded commercial FPGA-based network intrusion detection systems from companies including TippingPoint Technologies, Sensory Networks and Emulex Endace. Deep packet inspection has since evolved into deep content inspection, pioneered by Wedge Networks, which can do object level analysis of network traffic. As network speeds continue to increase and the importance of the Internet continues to rise, network filtering will increase in importance and is an ideal application of FPGA technology.

– Fast, Large-Scale String Match for a 10Gbps FPGA-Based Network Intrusion Detection System 2003 [Sourdis and Pnevmatikatos 2003].
This paper addressed the problem of pattern matching for intrusion detection which compares a set of strings against a (high speed) network traffic. It proposed a micro-architecture and a number of circuit-level optimizations exploiting parallelism between string matchers, pipelining of comparison units and large-fanout signals, parallel pipelined merge of individual results, etc. to achieve very high clock rate. The implementation on a Virtex2 FPGA exceeded 10Gbps processing throughput, a 3-5x improvement over previous work. Even ASICs approaches for NIDS pattern matching in the following years remained at this performance level [Tan and Sherwood 2005].
This paper paved the way for subsequent FPGA-based influential research that further exploited parallelism to improve pattern matching (area) efficiency while maintaining the 10Gbps performance goal such as [Clark and Schimmel 2004a] [Sourdis and Pnevmatikatos 2004]. Work in this area was later extended to cover regular expressions [Sourdis et al. 2008] [Brodie et al. 2006].

– Networks on Chip as Hardware Components of an OS for Reconfigurable Systems [Marescaux et al. 2003].
This paper is a significant refinement of a paper presented at FPL the previous year, which suggested that interconnection networks could enable fine-grained dynamic multi-tasking on FPGAs with low hardware overhead. The new paper proposed multiple networks on chip as hardware support for reconfigurable operating systems (OS), with reconfiguration, OS and application data communications being implemented as separate networks, each optimized appropriately, e.g. the reconfiguration network is optimized for latency and the application network optimized for bandwidth. This allowed the elegant support of features such as dynamic task relocation with state transfer, hardware debugging and security.
Subsequent operating systems supporting dynamic reconfiguration were inspired by ideas from this paper, notable examples including HERMES [Moraes et al. 2004], automotive runtime systems [Ullmann et al. 2004] and Dynoc [Bobda et al. 2005].

This quantitative study of hard networks-on-chip (NoC) in FPGAs described two novel implementations, mixed NoCs, which used hard routers and soft links; and hard NoCs, which utilized hard routers and hard links. When compared with soft NoCs implemented entirely using the FPGA fabric, the mixed/hard NoCs were 20x/23x smaller and 5x/6x faster. Moreover, a 64-node hard NoC adds less than 1 percent to the total area of a large FPGA. The authors found that hard NoCs consume 4.5-10.4 mJ of energy per GB of data transferred, this figure being comparable to the energy efficiency of soft point-to-point links requiring 4.7 mJ/GB.
The unexpected result that hard NoCs can be as energy efficient as the simplest traditional FPGA point-to-point soft interconnect, with minimal area overhead and higher flexibility makes a compelling case for including them in future devices. While the impact is difficult to judge at this early stage, this work has the potential to enhance the system-level interconnection of future FPGA applications.

Embedded NoCs have since been used to prototype Ethernet switching [Bitar et al. 2014] and packet processing [Bitar et al. 2015] applications on FPGAs, and the addition of embedded NoCs improved both the performance and efficiency of these applications significantly. Follow-on work in this area [Abdelfattah et al. 2015] focused on making embedded NoCs easier to use on FPGAs by defining design rules and creating CAD tools that enable hardware-friendly abstractions for this new kind of FPGA interconnect.

3. CONCLUSION

This paper highlighted a selection of publications which advanced the field over the last 25 years. We hope that it will help to inspire the next generation of researchers and industrialists over the next 25 years and beyond.

ACKNOWLEDGMENTS

This work was partially supported under Australian Research Councils Linkage Projects funding scheme (project number LP130101034) and Zomojo Pty Ltd.

REFERENCES


Lin Gan, Haohuan Fu, Wayne Luk, Chao Yang, Wei Xue, Xiaomeng Huang, Youhui Zhang, and Guangwen Yang. 2013. Accelerating solvers for global atmospheric equations through mixed-precision data flow engine. In Field Programmable Logic and Applications (FPL), 2013 23rd International Conference on. IEEE, 1–6. DOI: http://dx.doi.org/10.1109/FPL.2013.6645508


R. Glein, B. Schmidt, F. Rittner, J. Teich, and D. Ziener. 2014. A Self-Adaptive SEU Mitigation System for FPGAs with an Internal Block RAM Radiation Particle Sensor. In Field-Programmable Cus-
tom Computing Machines (FCCM), 2014 IEEE 22nd Annual International Symposium on. 251–258.
DOI: http://dx.doi.org/10.1109/FCCM.2014.79
Mike Hutton. 2015. Architectural Paths to Faster and More Robust FPGAs (Keynote). In Field-Programmable Technology (FPT), 2015 International Conference on.
Nachiket Kapre and André DeHon. 2011. VLIW-SCORE: Beyond C for sequential control of SPICE FPGA acceleration. In Field-Programmable Technology (FPT), 2011 International Conference on. 1–9. DOI: http://dx.doi.org/10.1109/FPT.2011.6132678

Received 23 Jan, 2016 (version 1); revised ; accepted