Performance and Energy Benefits of Instruction Set Extensions in an FPGA Soft Core *

Partha Biswas
partha@cecs.uci.edu

Sudarshan Banerjee
banerjee@cecs.uci.edu

Nikil Dutt
dutt@cecs.uci.edu

Center for Embedded Computer Systems
Donald Bren School of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425, USA

Paolo Ienne
paolo.ienne@epfl.ch

Laura Pozzi
laura.pozzi@epfl.ch

Ecole Polytechnique Fédérale de Lausanne (EPFL)
School of Computer and Communication Sciences
CH-1015 Lausanne, Switzerland

ABSTRACT

Performance of applications can be boosted by executing application-specific *Instruction Set Extensions (ISEs)* on a specialized hardware coupled with a processor core. Many commercially available customizable processors have communication overheads in their interface with the specialized hardware. However, existing ISE generation approaches have not considered customizable processors that have communication overheads at their interface. Furthermore, they have not characterized the energy benefits of such ISEs. We present a soft-processor customization framework that takes an input 'C' application and realizes a customized processor capturing the microarchitectural details of its interface with the specialized unit. We are able to accurately measure the speedup, energy, power and code size benefits of our ISE approach on a real system implementation by applying the design flow to a popular Xilinx Microblaze soft-processor core synthesized for four real-life applications. We show that only one large ISE per application is sufficient to get an average 1.41× speedup over pure software execution in spite of incurring communication overheads in the ISE implementation. We also observe a *simultaneous* savings in energy (up to 40%) and power (up to 12% peak power reduction) with this increased performance.

1. INTRODUCTION

Typically, applications running on a programmable platform can be executed either as a software algorithm or on a specialized hardware unit. The software approach is the slowest but most flexible while the hardware approach is the fastest but least flexible. *Instruction Set(IS)-extensible processors* comprise an emerging class of processors (especially in the embedded domain) that permit execution of only the critical application kernels in customized units (as hardware) with the rest of the application executing on the processor core (as software). This speeds up the application without compromising the processor clock or modifying the architectural model of the processor and yet preserves the flexibility of the software approach. We call such a coprocessing hardware element an *Ad-hoc Functional Unit (AFU)*. The AFU operation is triggered by an instruction or a set of instructions that we call an *Instruction Set Extension* or *ISE*. In the past, researchers have modeled AFUs having no communication overhead. However, many commercially popular customizable processors have communication overheads in their interface with AFUs. Therefore, our goal is to consider the microarchitectural details of an AFU interface in a processor customization framework and accurately evaluate the performance and energy benefits of ISEs in a realistic processor. The efficacy of the framework lies in seamlessly considering the synchronization between the processor and the AFU in a unified manner for different applications.

Minimizing power and energy consumption is as important as maximizing performance in embedded systems. A high power consumption may destroy a chip completely through overheating while a high energy consumption may reduce the battery life of an embedded device. Therefore, even though ISEs can achieve high speedups, designers need to determine if this speedup comes at a price of increased power. This paper shows that increased performance can also reduce both power and energy of a customizable processor in the presence of an AFU and reports the effects on code size and area.

It is predicted [17] that by 2010, over one-third of all PLD/FPGA devices are expected to have microprocessor cores, up from 15% today. Xilinx Microblaze [10] is a popular commercially-available soft-core. We demonstrate the use of our framework by transforming a given input application into a running Xilinx Microblaze hardware-software system. For four real-life applications (from Mediabench and EEMBC suites), we measure the real performance gain over pure software execution and also accurately evaluate energy and power consumption. Our experimental results show that significant speedup is obtained only when an ISE contains a large set of atomic operations. With only one large ISE per application, we obtained speedup of up to 1.47× over simple software execution and *simultaneously* up to 40% energy saving and 12% peak power reduction. To the best of our knowledge, this is also the first attempt to present the details of interfacing an AFU with a customizable soft-core. The main contributions highlighted in this paper are the following:

- We present a generalized interface-aware soft-processor customization framework for mapping an application in C into a running processor-AFU subsystem that enables

*This work was partially supported by NSF grants: CCR-0203813, CCR-0205712 and SRC contract: 2003-HJ1111.

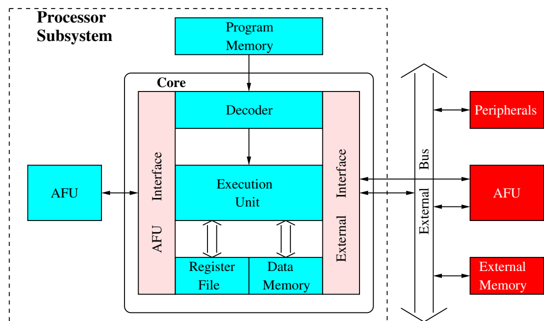

Figure 1: Target Customized Processor Subsystem

accurate evaluation of all the metrics deemed important in embedded system design, namely, performance, energy, power, cost and code size.

- By applying our framework to Microblaze soft-processor core, we conclude that ISEs can be simultaneously beneficial in terms of performance, energy, power and code size.

The rest of the paper is organized as follows. We present our target customizable processor model in Section 2. In Section 3, we present some related research work. We describe our framework for transforming a given application to a customized processor subsystem in Section 4. Section 5 presents how we use the framework to target Xilinx Microblaze soft-processor core. In Section 6, we describe our experimental results. Finally, Section 7 concludes the paper.

2. CUSTOMIZED PROCESSOR MODEL

Our goal is to map a given application to the target customizable processor model shown in Figure 1. In this model, the software part of the application stored in the program memory is composed of base instructions to be run on Execution Unit and ISEs to be run on the hardware part, i.e., AFUs. An AFU can be tightly-coupled with the core through an AFU interface inside the processor subsystem or loosely-coupled through an external bus. The AFU interface or the external interface implements the communication protocol between the AFU and the processor and thus controls synchronization of data and access to the processor register file.

The function of an ISE is to transfer control to an AFU for execution. An ISE can be either a **single user-defined** instruction or a set of **multiple pre-defined** instructions. A single user-defined instruction is decoded as a special instruction, which encapsulates inputs and outputs of an AFU as source and destination operands respectively. The decoder takes the responsibility of issuing such a special instruction to an appropriate AFU for execution. Alternatively, sending inputs and receiving outputs of the AFU from the processor can be done at the expense of multiple data transfer instructions. Such instructions must already exist in the instruction set of the processor in the form of “send data to AFU” and “receive data from AFU” instructions. In this case, the AFU incurs communication overhead at its interface while sending and receiving data.

In this paper, we present an integrated framework that drives a design flow from an application to a complete running system, which adheres to the customized processor model shown in Figure 1. We then apply our framework to a realistic soft-processor core and accurately study performance, energy, power, code size and cost of the implementation.

3. RELATED WORK

Several algorithms [1, 4, 2, 3, 5, 6] have recently been proposed to identify ISEs in a given application. The speedups over simple software execution claimed in most of the approaches [1, 4, 2, 3] are estimated by assuming a typical RISC processor execution model. The methodology in [5] targets Trimaran research infrastructure. Using a simulator, the authors show speedup for applications that reuse AFUs generated for other applications in the same domain. Such reuse of AFUs across application is possible only when ISEs found were reasonably small in size. However, we will confirm in our experimental results that such small-sized ISEs would not generate a considerable speedup for AFUs with communication overheads.

Sun et al. [6] employs a *Tensilica Instruction Extension (TIE)* compiler in their methodology and operates at a higher (C source-code) level of abstraction. Therefore, this methodology relies more on designer’s experience for ISE identification and mapping to AFUs. The AFU in this case therefore does not have any communication overhead. Fei et al. [7] integrated a fairly accurate energy estimation engine in the same framework, but they do not report a comparison of energy before and after extending the processor. A recent work having a goal of real system implementation [8] generated application-specific instructions for Altera Nios II processor in the presence of AFUs that do not have communication overheads. The results show a good speedup and limited area overhead, but they do not discuss energy or power consumption. Unlike [8], in this paper, we deal with the non-trivial details of synchronization between the processor and the AFU with the help of a generic communication template.

Note that in the prior related work, the AFU in general did not have communication overheads at its interface. Indeed, there are many commercially available processors providing such an interface. Common examples are Altera Nios II processor [13], LEON processor [12], etc. However, there are similarly many commercial customizable processors where AFUs incur overhead in sending and retrieving data. Some examples include STMicroelectronics ST120 [11], Xilinx Microblaze processor [10], etc. To the best of our knowledge, ISE generation in the context of AFUs incurring communication overheads at their interface with the core processor has not been studied yet. This is our motivation for proposing a framework that is capable of incorporating different AFU models and in particular, targeting Xilinx Microblaze soft-core. We apply the design flow of our framework to study performance gain, energy/power consumption, code size reduction and area overhead with the introduction of an AFU into the Microblaze subsystem.

4. OUR FRAMEWORK

Our framework takes as input a high-level application (in C), and generates an executable and an AFU with appropriate interfacing protocol (as shown in Figure 2).

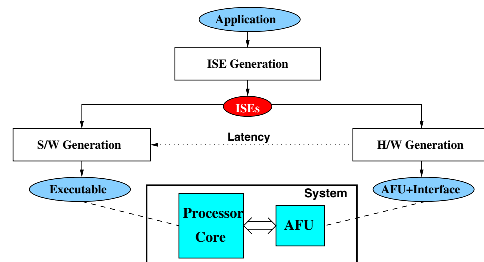

Figure 2: The Flow of our Framework

The expanded view of our framework is shown in Figure 3(a). It has three main phases: **ISE generation** phase, **S/W generation** phase, and **H/W generation** phase. The ISE generation phase generates ISEs under microarchitectural

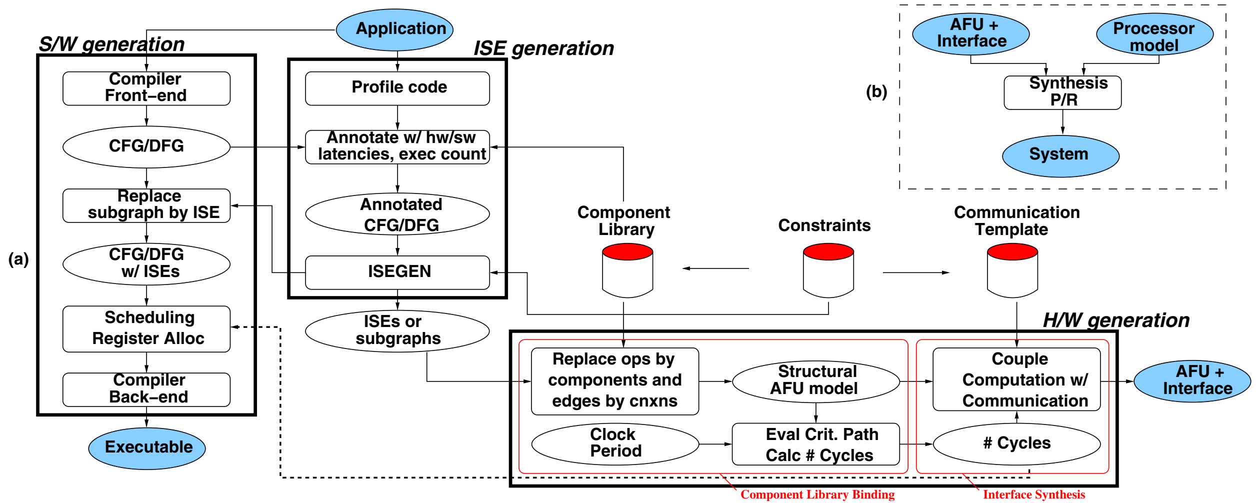

Figure 3: (a) A high-level application to a hardware/software system generation (b) Processor subsystem generation

constraints. The H/W generation phase synthesizes the corresponding AFUs with their interfaces and the S/W generation phase generates the executable. A dotted arrow between the two phases indicates that the latency of an ISE obtained in the H/W generation phase is passed on to the S/W generation phase. Finally, a post-processing phase builds the complete running system for evaluation.

4.1 Preprocessing Input Application

A compiler front-end yields *Control Flow Graph (CFG)* and *Data Flow Graph(DFG)* of an input application and runs predication to combine a set of small basic blocks into a large basic block. The input application is then profiled and the basic blocks are annotated with their execution counts. A **component library** is created containing a synthesizable combinational element corresponding to each instruction in the target instruction set. Each element in the library is synthesized for a given technology and the corresponding instruction in the *DFG* is annotated with a normalized hardware latency. Each instruction in the *DFG* is also annotated with its software latency obtained from the target architecture specification.

4.2 ISE Generation Phase

This phase is integrated with the compiler front-end. An ISE generation algorithm, e.g., [1, 4, 2, 3] takes the annotated *CFG/DFG* and returns subgraphs or ISEs that would maximize performance under microarchitectural constraints. The constraints are the maximum allowed number of AFUs, the maximum number of inputs/outputs and other microarchitectural feasibility constraints. For example, convexity constraint ensures that the dependency chain of instructions within an ISE is not intervened by any instruction outside the ISE. This guarantees that all the inputs of an ISE will be available at the time of executing the ISE. Although any ISE generation algorithm can be used, we use [1] in our framework because it identifies all the instances of an ISE exploiting large-scale ISE reuse.

4.3 H/W Generation Phase

The two subtasks of this phase are **component library binding** and **interface synthesis**. The identified subgraph or ISE is isolated and each instruction in the subgraph is replaced by the corresponding element in the component library. Figure 8 shows an example subgraph where each node maps to an element in the component library. The data dependencies between the instructions are replaced by port-to-port connections between the elements and the resulting structure is an AFU. This structural AFU model is then synthesized to evaluate the critical path length. The critical path length divided by the clock period of the processor core gives the number of cycles needed for the AFU operation. This latency information is passed on to the scheduler in the S/W generation phase (shown with a dotted arrow in Figure 3). The evaluated number of cycles is also used to synchronize the AFU with respect to the core.

Apart from the component library, the designer also creates a communication template for AFUs, which captures the communication protocol between the processor core and the AFU. The writing back of result from the AFU to the processor is delayed by the exact number of cycles required by the AFU operation. The implementation of communication protocol together with synchronization with the core completes the AFU interface synthesis. Note that the H/W generation phase can be applied to synthesize the AFU and its interface in the customized processor model presented in Figure 1.

4.4 S/W Generation Phase

This phase generates code for the target processor taking into account the presence of AFUs. The two subtasks in the S/W generation phase are **subgraph matching** and **subgraph replacement** with ISEs. Since all possible instances of an ISE have already been enumerated by the ISE generation phase, the subgraph matching simply consists of a DFG traversal and marking constituent instructions of the ISE in the DFG.

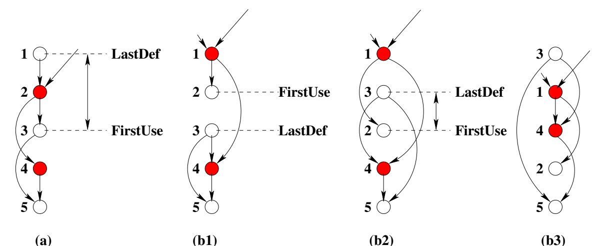

Figure 4: The ISE here is composed of the shaded instruction nodes. (a) An example showing the *LastDef* point and the *FirstUse* point; (b1) an example where it is not possible to insert the ISE under consideration; (b2) After code restructuring; (b3) positioning of the ISE between *LastDef* and *FirstUse*.

After subgraph matching, the ISE is used to replace the set of marked instructions in the DFG. We depict the ISE replacement strategy in Figure 4. An ISE can be placed any-

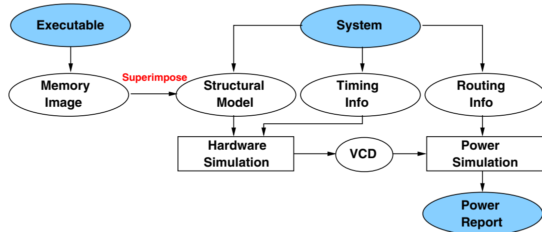

Figure 5: Measuring System Power

where between the point where its source operands have their last definition (*LastDef*) and the point where its destination operand has its first use (*FirstUse*) as shown in Figure 4(a) (the shaded nodes identify the ISE under consideration). Since ISE generation phase has ensured convexity of the identified subgraphs, it is never possible to have a dependency edge from the *FirstUse* node to the *LastDef* node (refer to [3] for the definition of convexity). Consider the following sequence of operations in instruction order: $(1)a = b * c; (2)f = a|0x2; (3)e = 5; (4)d = a + e; (5)g = e - d$. Suppose the ISE under consideration is a multiply followed by an add, as identified by the nodes labeled 1 and 4 in Figure 4(b1) respectively. Since in this case the *FirstUse* point appears earlier in the instruction chain than the *LastDef* point, the ISE cannot be placed anywhere (Figure 4(b1)). So, instruction reordering has to be done in order that the *LastDef* point precedes the *FirstUse* point. This reordering is possible because there is no dependency from *FirstUse* to *LastDef*. Figure 4(b2) shows the code snippet after restructuring Figure 4(b1) (i.e., swapping the positions of node 2 and node 3) and Figure 4(b3) shows the placement of ISE between the *LastDef* point (node 3) and the *FirstUse* point (node 2).

If an ISE is used as a single user-defined instruction, a single instruction just replaces the set of constituent instructions. Replacing the multiply and the add with a single user-defined instruction ($ISE1(\cdot,\cdot,\cdot)$), the resulting instruction sequence (as in Figure 4(b3)) would become: (3) $e = 5$; (1),(4) $d = ISE1(b, c, e)$; (2) $f = a|0x2$; (5) $g = e - d$. However, if an ISE is represented as a set of predefined data transfer instructions ($send(\cdot)$, $receive(\cdot)$), the resulting instruction sequence after ISE replacement would appear as: (3) $e = 5$; (1),(4) $send(b)$; $send(c)$; $send(e)$; $receive(d)$; (2) $f = a|0x2$; (5) $g = e - d$. After subgraph replacement with ISE, the compiler performs scheduling, register allocation and target code generation as a back-end pass. Note that the latency of the ISE required by the scheduler is derived from the H/W generation phase as shown in Figure 3(a).

4.5 Targeting a Soft-core

As a final step, the processor model of the target Soft-core along with the AFU and its interface are synthesized and implemented using standard synthesis and Place-and-Route tools (Figure 3(b)). The executable generated in Figure 3(a) and the system synthesized in Figure 3(b) are deployed in two schemes, one for measuring speedup and the other for evaluating energy/power consumption. With the goal of measuring actual time spent in running the application, the **scheme for Performance Measurement** uses the bitmap of the synthesized system to program an FPGA fabric, which then becomes the platform for actually running the executable. The executable is downloaded into the system memory through a JTAG port and the number of cycles for running the executable is measured using a hardware timer.

Since there is no direct way to measure power of a running system on the FPGA fabric, we employ a different **scheme for Power/Energy Evaluation** (depicted in Figure 5) for accurately evaluating the power and energy consumption of the system. Note that there are three kinds of information in the post-Place-and-Route system (Figure 3(b)): the structural model of the system, the timing information and the routing information. We superimpose the memory image of the executable (in Figure 3(a)) into the memory section of the structural model. This complete structural model along with the timing information is run through a cycle-accurate hardware simulator to generate a *Value Change Dump (VCD)* of all the signals in the structural netlist. The routing information and the VCD information together are then used by a power simulator to generate the dynamic power consumed at different time steps. We then derive the total energy dissipated in the system from the reported power and the measured execution time.

5. TARGETING MICROBLAZE

We illustrate the utility of our framework on a popular representative platform: Xilinx Microblaze [10], an IS-extensible soft processor that allows an AFU to be connected with the processor via *Fast Simplex Links (or FSLs)*. FSLs are dedicated point-to-point unidirectional 32-bit wide FIFO interfaces. The Microblaze is capable of including a maximum of 8 input and 8 output FSLs.

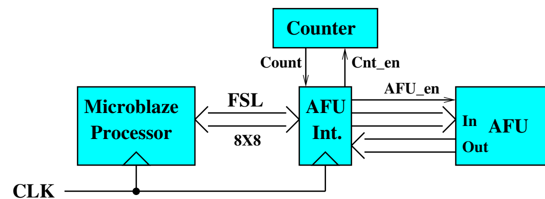

Figure 6: Microblaze Processor Core with an AFU and its Interface.

Microblaze is a 32-bit RISC processor with a simple 3-stage pipeline. Figure 6 shows an AFU and its interfacing with the Microblaze processor core via 8×8 FSL channels. The AFU interface implements the processor-AFU communication protocol and is synchronous with the Microblaze processor through a global clock (*CLK*). The AFU interface is also connected to a counter module to enable counting whenever required. If the count enable signal (*Cnt_en*) is '1', counting is enabled. Otherwise, the counter is reset to '0'. The signals *In[32]* and *Out[32]* are used to send data to and receive data from the AFU respectively. When the AFU-enable signal, *AFU_en* is '1', the AFU latches the output in *Out[32]*.

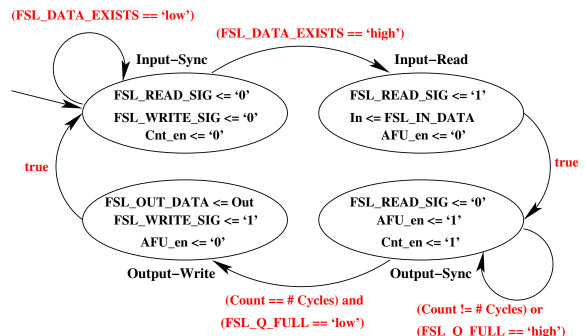

Figure 7: Communication Template for AFU Interface in Microblaze

In Figure 7, we present the generic communication template for Microblaze-AFU interaction as a *Finite State Machine (FSM)* synchronous with respect to *CLK*. For the sake of explanation, we call an FSL channel FSL_R when it is used for AFU read operation or FSL_W when it is used for AFU write operation. Associated with every FSL_R

channel is a set of three signals, namely, (*FSL_READ_SIG*, *FSL_DATA_EXISTS*, *FSL_IN_DATA[32]*). Another triplet, (*FSL_WRITE_SIG*, *FSL_FIFO_FULL*, *FSL_OUT_DATA[32]*) is associated with every FSL_W channel. The FSM is initially in "Input_Sync" state waiting for data to arrive on an FSL_R channel. When data exists on the FSL channel, the corresponding *FSL_DATA_EXISTS* signal goes high causing a transition from "Input_Sync" state to "Input_Read" state. In "Input_Read" state, *FSL_READ_SIG* is set to high to cause the data in the FSL_R FIFO to be read into *In[32]* using a 32-bit signal array, *FSL_IN_DATA*. After the data has been read into *In[32]*, the FSM transitions to "Output_Sync" state and waits on the AFU operation by enabling the counter. After *# Cycles* (as evaluated in the H/W generation phase in Figure 3(a)) has elapsed, the result of the AFU operation is latched in *Out[32]*. If FSL_W FIFO is not full (i.e., *FSL_FIFO_FULL* is low), a state transition takes place to "Output_Write" state. In the "Output_Write" state, data from *Out[32]* is written into the FSL_W FIFO using *FSL_OUT_DATA[32]* by setting *FSL_WRITE_SIG* to high. Thus, for introducing every new AFU, only the AFU module in Figure 6 and the *# Cycles* change in the process of H/W generation, while the communication template is reused.

6. EXPERIMENTS

We demonstrate the effectiveness of our approach using a number of front-end tools in our framework shown in Figure 3(a).

6.1 Experimental Setup

The ISE generation algorithm [1] was integrated with a Machine SUIF [9] front-end. The S/W generation was done with Microblaze GCC-2.95 (*mb-gcc*) compiler. Microblaze Instruction Set has multiple data-transfer instructions for sending data to and receiving data from its FSL channels — *put* for sending and *get* for receiving data in blocking mode, and *nput*/*nget* are the corresponding instructions in non-blocking mode. We used the non-blocking send instruction (*nput*) and the blocking receive instruction (*get*) for our AFU interface. Because of using two different compilers for ISE generation and S/W generation, the subgraph replacement with ISEs was done as a post-assembly pass on the assembly output of *mb-gcc*. After replacing the identified subgraphs with ISEs, *mb-gcc* was run again to generate the executable.

We selected four real-life applications for demonstrating the effectiveness of our framework: *autcor* (Auto-correlation) from EEMBC suite, *adpcm-e* (ADPCM Encoder) and *adpcm-d* (ADPCM Decoder) from Mediabench suite, and *AES* (AES encryption). Our platform is *Xilinx Multimedia Board*, which is equipped with a *Virtex-II XC2V2000* FPGA. We used *Xilinx Platform Studio* for configuring the FPGA to include a Microblaze processor with a 64KB (i.e., the maximum size possible) Block RAM (BRAM), two Local Memory Buses (LMBs) (to interface with BRAM – one for instruction and the other for data), one Microblaze Debugging Manager (MDM) and one Timer (both MDM and Timer on a single On-chip Peripheral Bus (OPB)). The standard inputs and outputs of an application were redirected to the MDM and the elapsed number of cycles was evaluated using the Timer. We set the clock frequency of the Microblaze processor to 50 MHz. The tools used in the second scheme (Figure 5) for evaluating energy and power are ModelSim for hardware simulation [15] and Xilinx XPower for power simulation [16].

6.2 Performance and Code Size

The code generation for the baseline configuration was done by *mb-gcc* with all optimizations turned on (*-O2*, *-mnoxl-soft-mul*) so that the performance is maximized in pure software execution. The Microblaze configuration was then customized for different applications by introducing AFU with its interface as explained in Section 5. The ISEs were generated with I/O constraints of maximum 4 inputs and 2 outputs and number of AFUs set to 1. Note here that for each application, a different Microblaze configuration is generated and the resulting system is analyzed by applying our framework. The results in terms of code size reduction and speedup over software execution are summarized in Table 1.

Table 1: Speedup and Code Size Reduction with the Introduction of an AFU having 4 inputs and 2 outputs in the Microblaze subsystem

BMs	Core Only		Core + AFU		Code	Spdup
	Bytes	Cycles	Bytes	Cycles	Redn	
autcor	58444	264305	58452	404673	-8	0.65×
adpcm-d	12049	252688	11953	190979	96	1.32×
adpcm-e	14121	157177	13989	106821	132	1.47×
AES	16013	240613	14957	167397	1056	1.44×

Each of the operand-send and result-receive operations in Microblaze has a latency of 2 cycles. Consequently, the latency for transferring 6 operands is 12 cycles in the worst case and 6 cycles in the best case (i.e., if all the latencies are successfully hidden by the scheduler). The ISE generated for *autcor* was a chain of just three operations: a multiply, a barrel right shift and an add having software latencies as 3, 2 and 1 cycles respectively. With AFU operation taking just 1 cycle, the best case latency of the ISE is $6+1=7$ cycles. Thus, even the best case performance of the ISE lags behind the worst case performance of the corresponding software execution ($3+2+1=6$ cycles). Consequently, there was slowdown instead of speedup for *autcor* owing to the communication overhead. However, there are some prior related work [6, 8], which have shown speedup even with small-sized ISEs containing on the order of 3-4 instructions because of incurring no communication overhead in processor-AFU interface. **Thus, we confirm that if the AFU interface has a communication overhead, a small-sized ISE will only result in performance degradation.**

The applications *adpcm-d* and *adpcm-e* are the two examples where predication of several small critical basic blocks led to a large basic block. Consequently, the ISEs found for these two benchmarks are very large containing on the order of 40 operations. This led to a significant speedup in spite of the communication overhead. Figure 8 shows the ISE of *adpcm-e* that generated a speedup of 1.47× over pure software execution. The shaded nodes show the inputs and the outputs of the ISE.

Table 2: Power Benefits of ISEs in the Microblaze subsystem

BMs	Core Only		Core + AFU		% Pk Pwr Redn	% Avg Pwr Redn
	P. Pwr (mW)	A. Pwr (mW)	P. Pwr (mW)	A. Pwr (mW)		
autcor	1957	1287	1869	1229	4.5	4.5
adpcm-d	1975	1317	1919	1197	2.8	9.1
adpcm-e	2070	1332	2012	1178	2.8	11.6
AES	2256	1276	1982	1187	12.1	7.0

The last benchmark under consideration is *AES*, which has the largest number of instructions in its critical basic block. The generated ISE [1] had 8 instances in the critical basic block covering more than 50% of the DFG and overall 12 instances in the critical function. Both the large size and large-scale reuse (as defined in [1]) of the ISE accounts for a significant speedup (1.44×) obtained on AES despite the overhead in sending and receiving operands. Along with the merit of speedup, *AES* also exhibit a 7% code size reduction owing to replacement of a large chunk of code by an ISE in the form of a set of data transfer instructions.

Figure 8: **An ISE for ADPCM ENCODER (*adpcm-e*) having 4 inputs and 2 outputs; each operation node maps to a hardware component.**

6.3 Power and Energy Results

From Table 2, it is evident that both the peak power (P. Pwr) as well as the average power (A. Pwr) reduced with the introduction of AFU. Because the presence of both core and AFU apparently indicates more circuit activity, an initial expectation is increased power with the addition of AFU. However, because the ISE here is a multi-cycle operation interlocked with the Microblaze pipeline, the AFU operation completely overlaps with a processor pipeline stall. Consequently, we obtain an overall power reduction in the presence of AFU operation owing to reduced overall circuit activity.

Table 3: Energy Benefits of ISEs in the Microblaze subsystem

BMs	Tot Energy (μJ) for Core Only	Tot Energy (μJ) for Core+AFU	%age Saving	Spdup
autcor	2.21	3.10	-40.27	0.65×
adpcm-d	8.48	5.84	31.13	1.32×
adpcm-e	10.54	6.34	39.85	1.47×
AES	69.09	43.69	36.76	1.44×

As shown in Table 3, we also obtained up to 40% saving in energy on account of reduced application runtime. It is interesting to note that the trend of energy decrease (or increase) exactly follows that of speedup (shown again in Table 3 for the sake of comparison). This trend can be expected as a corollary to a consistent power reduction shown in Table 2. **Thus, contrary to conventional expectation, enhanced performance *simultaneously* results in reduced power and energy for the customized Microblaze soft-core.**

6.4 Slices Utilization

The *XC2V2000* FPGA that we use as our target platform has 10752 slices. Table 4 shows the percentage utilization of the FPGA slices before and after introducing the AFU that brought the speedup in Table 1.

Note here that *XC2V2000* used here is very small. The largest possible *Virtex-II* chip, *XC2V8000* contains 46592 slices. If the largest FPGA is used instead of *XC2V2000*, the

Table 4: Slices Utilization (out of 10752) in the absence of an AFU and in the presence of an AFU for the four applications in XC2V2000 FPGA

BMs	No AFU	autcor	adpcm-d	adpcm-e	AES
Slices	1274	1609	1804	2226	2043
Util.	11%	14%	16%	20%	19%

average slices utilization reduces to only 5%, which is very reasonable. **Thus, the area overhead of including an AFU in the Microblaze subsystem is also minimal.**

7. SUMMARY AND FUTURE DIRECTIONS

Applications can be accelerated in a programmable processor by executing their performance-critical sections in customized *Ad-hoc Functional Units (AFUs)* as *Instruction Set Extensions (ISEs)*. We presented an interface-aware processor customization framework that enabled us to implement a customizable soft-core microarchitecture capturing the details of interfacing with an AFU. We applied our framework to four real-life applications and realized four different processor configurations. Our results confirmed that in the presence of communication overhead at the processor-AFU interface, significant speedup over pure software execution is possible only if the AFU function is sufficiently larger than a set of 2-3 operations. Further analysis of the synthesized systems led to the conclusion that integration of AFUs in a customizable processor can result in increased performance and reduced code size, while *simultaneously* decreasing power and energy consumption. Our future work will investigate the advantages of ISEs in other reconfigurable platforms and commercially available processors.

8. REFERENCES

[1] P. Biswas, S. Banerjee, N. Dutt, L. Pozzi and P. Ienne. ISEGEN: Generation of High-Quality Instruction Set Extensions by Iterative Improvement. In *Proc. of DATE*, 2005.

[2] P. Biswas, V. Choudhary, K. Atasu, L. Pozzi, P. Ienne and N. Dutt. Introduction of Local Memory Elements in Instruction Set Extensions. In *Proc. of DAC*, 2004.

[3] K. Atasu, L. Pozzi and P. Ienne. Automatic Application-Specific Instruction-Set Extensions under Microarchitectural Constraints. In *Proc. of DAC*, 2003.

[4] P. Yu and T. Mitra. Scalable Custom Instructions Identification for Instruction-Set Extensible Processors. In *Proc. of CASES*, 2004.

[5] N. Clark, H. Zhong and S. Mahlke. Processor Acceleration through Automated Instruction Set Customization. In *Proc. of MICRO*, 2003.

[6] F. Sun, S. Ravi, A. Raghunathan and N. K. Jha. Synthesis of Custom Processors based on Extensible Platforms. In *Proc. of ICCAD*, 2002.

[7] F. Sun, S. Ravi, A. Raghunathan and N. K. Jha. A Hybrid Energy-Estimation Technique for Extensible Processors. *IEEE TCAD*, 2004.

[8] J. Cong, Y. Fan, G. Han and Z. Zhang. Application-Specific Instruction Generation for Configurable Processor Architectures. In *Proc. of FPGA*, 2004.

[9] Machine SUIF. http://www.eecs.harvard.edu/hube/software/software.html.

[10] Microblaze Processor Reference Guide. http://www.xilinx.com/ise/embedded/mb_ref_guide.pdf.

[11] ST100 DSP Core Architecture Overview. http://www.st.com/stonline/prodpres/dedicate/st100/overview/overview.htm.

[12] The Leon Processor User Manual. http://www.ra.informatik.uni-stuttgart.de/~virazela/LP_Project/leon-2.3.7.pdf.

[13] The Nios II Processor Reference Handbook. http://www.altera.com/literature/hb/nios2/n2cpu_nii5v1.pdf

[14] SC140 DSP Core Reference Manual. http://www.soc.napier.ac.uk/module.php3?op=getresource&cloaking=no&resourceid=1473119.

[15] ModelSim SE datasheet http://www.model.com/products/pdf/datasheets/se.pdf

[16] Xilinx XPower Documentation http://toolbox.xilinx.com/docsan/xilinx6/books/data/docs/dev/dev0089_14.html

[17] Panelists peer into future of FPGAs. Article 60407325, *EETimes*. March 7, 2005.