Appeared in *IET Computers & Digital Techniques*, Vol. 6, Issue 4, ISSN: 1751-8601, pp. 205-213, 2012

Multi-Objective Optimizations for a Superscalar Architecture with Selective Value Prediction

Arpad Gellert, Horia Calborean, Lucian Vintan, Adrian Florea

Computer Engineering Department, “Lucian Blaga” University of Sibiu, Emil Cioran Street, No. 4, 550025 Sibiu, Romania
E-mail: {arpad.gellert, horia.calborean, lucian.vintan, adrian.florea}@ulbsibiu.ro

Abstract

This work extends an earlier manual design space exploration of our developed Selective Load Value Prediction based superscalar architecture to the L2 unified cache. After that we perform an automatic design space exploration using a special developed software tool by varying several architectural parameters. Our goal is to find optimal configurations in terms of CPI (Cycles per Instruction) and energy consumption. By varying 19 architectural parameters, as we proposed, the design space is over 2.5 millions of billions configurations which obviously means that only heuristic search can be considered. Therefore, we propose different methods of automatic design space exploration based on our developed FADSE tool which allow us to evaluate only 2500 configurations of the above mentioned huge design space! The experimental results show that our automatic design space exploration (DSE) provides significantly better configurations than our previous manual DSE approach, considering the proposed multi-objective approach.

Keywords: Multi-Objective Design Space Exploration, NSGA-II, Simulation, Superscalar, Value Prediction, Cache

1. Introduction

This paper presents an automatic design space exploration of an architecture containing a Selective Load Value Prediction scheme suitable for energy-aware superscalar processors. A load value predictor is a hardware architectural enhancement which speculates over the results of load instructions to speed-up the execution of the subsequent instructions. Lipasti first proposed the Load Value Prediction (LVP) concept and, particularly, he developed a non-selective LVP in 1996 but we haven’t used his LVP structure due to its huge complexity. Our proposed architectural enhancement (SLVP) differs from a classic value predictor due to an improved selection scheme that allows activating the predictor only when a miss occurs in the first level of cache. We use a simple direct-mapped table which requires less additional hardware enabling reduced energy consumption than traditional approaches (Lipasti’s first proposal).

We first used our SLVP in [9] where we have analyzed the efficiency of selectively anticipating the results of long-latency instructions within superscalar and Simultaneous Multithreaded (SMT) architectures. Particularly we have focused on Multiply, Division and critical Loads (with miss in L1 data cache). We integrated into the M-SIM simulator

[22] a Dynamic Instruction Reuse scheme for the Mul/Div instructions and a Last Value Predictor for the critical Load instructions. Our improved superscalar architecture achieved an average Instruction Per Cycle (IPC) speedup of 3.5% on the integer SPEC 2000 benchmarks, of 23.6% on the floating-point benchmarks, and an improvement in energy-delay product of 6.2% and 34.5%, respectively. Our evaluations have also shown higher IPC and lower relative energy consumption (energy-delay product) on all the evaluated SMT configurations (1, 2, 3 and 6 threads).

Since our previous results show that most of the IPC speedup was generated by the Load Value Predictor we further focalized on this speculative technique. In [10] we performed a manual design space exploration regarding the size of the L1 data cache in order to find the optimal configuration, which keeps high performance at low energy consumption. We have shown that the performance lost by reducing the L1 cache capacity can be covered by our Selective Load Value Prediction (SLVP) technique. The experimental results, performed on the SPEC 2000 benchmarks, have expressed that reducing the L1 data cache space by quartering its size and using SLVP produces an improvement of the IPC and energy consumption in both the superscalar and SMT architectures against the corresponding baseline architectures.

In this work we extend the manual design space exploration of a SLVP-based superscalar architecture to the L2 unified cache. Our goal is to find optimal configurations in terms of CPI (Cycles per Instruction) and energy consumption. After these manual design space exploration performed by varying only 2 parameters we intend to increase the number of the varied parameters to 19. By varying 19 architectural parameters, as we proposed, the design space is over 2.5 millions of billions configurations which obviously means that only heuristic search can be considered. Therefore, we propose different methods of automatic design space exploration based on our developed FADSE tool which allow us to evaluate around 2500 configurations of the above mentioned huge design space, while still finding good solutions! We implemented a domain ontology consisting of some micro-architectural restrictions and expert knowledge expressed through fuzzy rules, in order to accelerate the design space exploration. By performing a multi-objective automatic design space exploration of the same architecture (using our developed FADSE tool [3]), but varying 19 architectural parameters, the obtained configurations are significantly better than the manually obtained ones.

The paper is organized as follows. Section 2 reviews the state-of-the art of value prediction techniques and some basic concepts about design space exploration. Section 3 introduces the target architecture and presents the simulation methodology. Also, there is performed a short presentation of the used metrics. Sections 4 and 5 describe the manual and automatic design space exploration, respectively, together with experimental results obtained on the Alpha AXP 21264 architecture. Finally, Section 6 summarizes the relevant contributions of this work and presents some further work directions.

2. Related Work

Lipasti et al. [16] originally introduced the Load Value Prediction as a new data-speculative micro-architectural technique exploiting the concept of value locality and the dynamic correlation between load instruction address and its actual value. An important difference between our value prediction approach and Lipasti's is that we selectively predict only the Load instructions that generate a miss in L1 cache. Thus, we attenuate the mispredictions cost and reduce the hardware cost of the speculative micro-architecture. Moreover, since less hardware is required, there is also less power consumption.

Other value predictors like the stride-, context- and perceptron-based, have been proposed in our earlier work [26] for register centric value prediction. The SLVP was already used in our previous papers [9, 10] and other load value predictors have been proposed in [4, 19]. Different architectural support techniques for value prediction or energy efficient approaches that speculate over the results of load instructions in order to speed-up the execution are presented in [21, 11]. In [21] the authors selectively re-execute only the speculatively-retired instructions that depended on the mispredicted value, so called ForwardSlice, performing a speed-up execution with an additional hardware budget (ReSlice buffer). Somewhat similar, in [11] the author proposes a new load latency tolerant design that is both energy efficient, and applicable to both in-order and out-of-order cores, based on slice re-execution as an alternative use of multi-threading support, efficient schemes for register and memory state management, using a chained store buffer for efficient store-to-load forwarding, and using pruning mechanisms to reduce re-execution overheads. The main idea of load latency tolerance is to virtually scale the critical execution structures (issue queue, physical register file). Load latency tolerance designs remove from these window resources all dependent instructions of loads with miss in caches in order to allow younger instructions to enter the pipeline and execute. When a miss returns, the instructions which depend on it are re-injected into the pipeline – re-acquiring issue queue entries and physical registers – and re-execute.

Further in this section we present some of the best known design space exploration tools. M3Explorer [25] is a DSE framework that includes many design space exploration algorithms. M3Explorer can use response surface models to accelerate the design space exploration. Another DSE tool is in a form of a website: archexplorer.org [7]. The users can upload their component on the website where it is integrated into a computer system simulator. The design is compared against other designs introduced by other users. The users do not have any control on the algorithm being used. NASA [14] is also a similar tool. It allows the user to easily integrate his/hers DSE algorithm and offers the possibility to connect to any simulator (features offered also by FADSE). Magellan [15] is a DSE tool which is bounded to a certain simulator (SMTSIM). Magellan can perform only single objective DSEs.

In [12] we used our developed FADSE tool to explore the vast design space of the Grid Alu Processor (GAP) and its post-link optimizer called GAPtimize, both developed at Augsburg University. It has shown that FADSE is able to thoroughly explore the design space for both GAP and GAPtimize and it can find an approximation of the Pareto [2, 3] frontier consisting of near-optimal individuals in moderate time. For the GAP, FADSE can find, due to the approximation of the complexity, efficient configurations.

To our knowledge FADSE is the single DSE tool that allows the user to introduce domain knowledge through fuzzy rules, written in a human-readable form, in order to accelerate the design space exploration. Mariani et al. [18] use neural networks to accelerate the DSE process. The authors predict through neural networks if an individual is worth simulating or not, but still the knowledge of an architect is not used.

3. Simulation Methodology

All the experimental results presented further were obtained using SPEC 2000 benchmarks on 500 million dynamic instructions, skipping the first 300 million instructions. We evaluated six floating-point benchmarks (*applu*, *equake*, *galgel*, *lucas*, *mesa*, *mgrid*) and six integer benchmarks: computation intensive (*bzip*, *gcc*, *gzip*) and memory intensive (*mcf*, *twolf*, *vpr*). Our measurements are generated using an 80 nm CMOS technology and 1.2 GHz frequency.

The target architecture is a superscalar Alpha AXP 21264 processor augmented with a direct mapped Selective Load Value Predictor of 1024 entries, access latency of 1 cycle and prediction latency of 3 cycles [10]. It has a Register File of [32 int / 32 fp]*8, a Reorder Buffer (ROB) of 128 entries and a Load/Store Queue (LSQ) of 48 entries. First-level caches are 64 KB, 2-way associative, with a 1-cycle latency. The second-level unified cache is 4 MB, 8-way associative and 6-cycle latency. The main memory has a latency of 100 cycles.

For the performance metrics we chose CPI (and not IPC) because we want to minimize all the objectives for the clarity of the Pareto graphs. For the relative CPI reduction we used the following formula:

$$CPI_{reduction} = \frac{CPI_{base} - CPI_{improved}}{CPI_{base}} \cdot 100\ [\%] \qquad (1)$$

where CPI_{base} and $CPI_{improved}$ are cycles per instructions with the baseline and improved architectures, respectively. A positive value of $CPI_{reduction}$ means a performance improvement related to the baseline architecture.

The detailed power modeling methodology, used in the simulator, is presented in [1]. The dynamic power consumption in CMOS microprocessors is defined as:

$$P = C \cdot V_{dd}^2 \cdot a \cdot f \qquad (2)$$

where *C* is the capacitance, generated using *Cacti* [23], V_{dd} is the supply voltage, and *f* is the clock frequency. V_{dd} and *f* depend on the assumed process technology. The activity factor *a* indicates how often clock ticks lead to switching activity on average. The power consumption of the modeled units highly depends on the internal capacitances of the circuits. From the capacitance point of view, there are three categories of architectural structures: array structures, content-associate memories, and complex logic blocks. The first two categories are used to model the caches, branch predictors, the reorder buffer, the register renaming table, and the register file, while the last category is used to model functional units.

For the power consumption evaluation we used the *aggressive non-ideal conditional clocking* model [5] which scales linearly the power of active units with their usage and assumes 10% power dissipation in the case of unused units. The instantaneous average power consumption (P_{Mean}) for a certain benchmark is computed with the following relation:

$$P_{Mean} = \frac{\int_{0}^{T} P(t) \cdot dt}{T} \tag{3}$$

where *T* is the total simulation time in cycles and *P* is given in relation (2). The energy consumption is given by:

$$E = P_{Mean} \cdot T \tag{4}$$

where P_{Mean} is computed with relation (3). The average energy (weighted mean) is given by the following formula in $[W \cdot cycles]$:

$$E_{Mean} = \frac{\sum_{i=1}^{N} E_i \cdot T_i}{\sum_{i=1}^{N} T_i} \tag{5}$$

where *N* is the number of benchmarks, E_i is the total or per unit energy computed for benchmark *i* and T_i is the total number of cycles executed within benchmark *i*. The energy reduction percentage is given by:

$$E_{reduction} = \frac{E_{base} - E_{improved}}{E_{base}} \cdot 100\,[\%] \tag{6}$$

where, E_{base} and $E_{improved}$ are the energy consumptions of the baseline and our improved architectures, respectively. Thus, a positive value of $E_{reduction}$ means an improvement of the relative energy consumption.

4. Manual design space exploration of the unified L2 cache

A method to increase the cache performance is to reduce the penalty in case of miss using multilevel caches. Our simulated architecture uses two level exclusive caches. This allows smaller L2 data caches involving less power consumption. The evictions are performed based on the Least Recently Used (LRU) algorithm for both cache levels.

The first goal of our research consists in performing a design space exploration regarding the sizes of the L1 data cache and the unified (instruction & data) L2 cache in superscalar architectures augmented with SLVP structures. Thus, we will double, halve, quarter and eighth the L2 cache and we will halve, quarter and eighth the L1 data cache, considering as reference the architecture presented in Section 3. We note with *m*UL2_*n*DL1 a configuration using m*4 MB 8-way associative unified L2 cache (m=2, 1, 1/2, 1/4, 1/8) and n*64 KB 2-way associative L1 data cache (n=1, 1/2, 1/4, 1/8).

Figure 1 presents the relative CPI and energy reduction – computed based on formulas (1) and (2), respectively – of different configurations with SLVP of 1024 entries reported to the baseline configuration without SLVP. It can be observed that the SLVP helps maintaining a better CPI and energy consumption when the cache sizes are reduced. The CPI reduction with the help of the SLVP is positive up to using halve of UL2 and eighth of DL1. Starting with quartering UL2, the CPI reduction is negative; therefore no performance improvement is achieved.

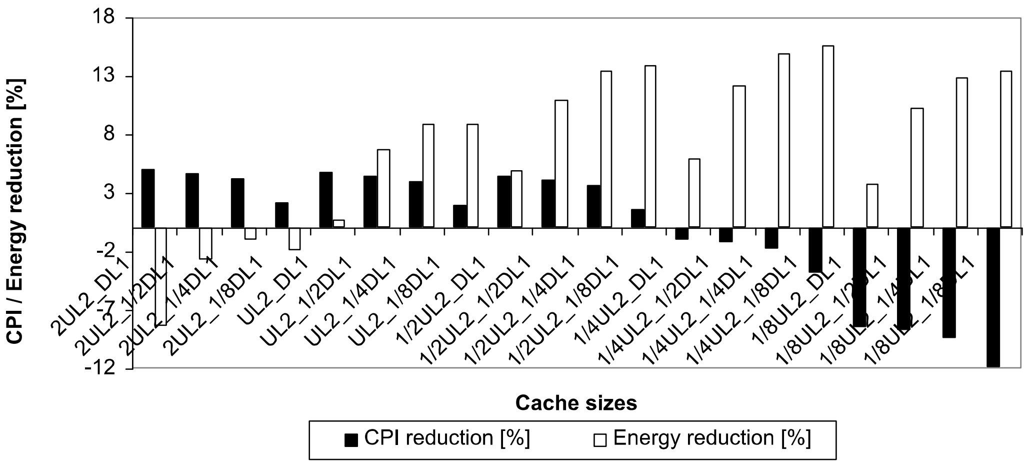

Figure 1. Relative CPI and energy reduction reported to UL2_DL1 without SLVP as baseline

The energy reduction is lower in the case of reducing the L2 cache to 1/8 than in the case of quartering it. The energy consumption has a static and a dynamic component. 1/8UL2 cache implies a higher miss rate than 1/4UL2 and therefore higher dynamic power consumption, due to the higher number of off-chip accesses. Thus, even if the static power consumption of 1/8UL2 is lower than of 1/4UL2 the energy consumption is higher due to the higher dynamic power consumption. Therefore, using only the quarter of the L2 cache (2 MB) and the eighth of the L1 data cache (8 KB) is optimal from the energy consumption viewpoint.

As a preliminary conclusion, after the manual design space exploration the best configuration regarding CPI is 2UL2_DL1 whereas the best configuration in terms of energy consumption is 1/4UL2_1/8DL1. There are also some optimal configurations from both CPI and energy viewpoints: 1/2UL2_1/2DL1 and 1/2UL2_1/4DL1. These results obtained through manual DSE encourage us to explore a larger design space by automatic DSE because the best and the optimal configurations are different and there are also other parameters which can be varied.

5. Automatic design space exploration

In the previous section we varied only the cache sizes through our manual design space exploration. Beside the caches there are several parameters that can highly influence our two objectives: CPI and energy consumption. We selected 19 important architectural parameters to be varied during our automatic design space exploration, with the lower and upper limits given in Table 1. By varying these 19 architectural parameters the design space grows over $2.5*10^{15}$ (2.5 millions of billions) configurations which obviously means that only heuristic search can be considered. Therefore, we propose different methods of automatic design space exploration based on our developed FADSE tool that contains also a NSGA-II genetic algorithm implementation.

Parameter		Lower limit	Upper limit
DL1 / IL1 cache	Sets	2	32768
	Block size (bytes)	8	256
	Associativity	1	8
UL2 cache	Sets	256	2097152
	Block size (bytes)	64	256
	Associativity	2	16
SLVP (entries)		16	8192
Decode / Issue / Commit width		2	32
ROB / LSQ / IQ size (entries)		32	1024
Number of physical register sets (int / fp)		2/2	8/8
Int / fp ALU		2	8
Int / fp MUL/DIV		1	8

Table 1. Parameter limits

To perform design space exploration we have developed a tool called Framework for Automatic Design Space Exploration (FADSE). It includes many state of the art evolutionary algorithms through the included jMetal [8] library.

FADSE can be connected to almost any existing simulator. The parameters are described through an extensible XML interface. FADSE allows parallel evaluation (included algorithms had to be modified to allow this).

FADSE is a client-server application. The number of clients can be dynamically changed. Clients can be stopped or started while the DSE process runs. Since performing DSE can take a lot of time (weeks), reliability of the DSE tool is a major concern. FADSE is able to cope with failing clients, failing networks or even power loss of the entire system. It is able to recover from these situations by detecting the problems and resubmitting the simulations to other clients. In case of power loss, it can restart the DSE process by making use of the integrated checkpointing mechanism. It contains a database which allows reusing already simulated individuals. This leads to a reduction of the time required to perform an exploration process. FADSE includes many metrics that the user can choose to evaluate the DSE process or to compare different algorithms. Some of the implemented metrics are: hypervolume, coverage, two set difference hypervolume [24], etc.

We have chosen for our automatic DSE the NSGA-II genetic algorithm. NSGA-II is a multi-objective genetic algorithm developed by Deb et al. [6]. NSGA-II has been chosen as it provided very good results in our previous experiments. In [2] we showed that NSGA-II, SMPSO [20] and SPEA2 [28] obtain results of similar quality but SPEA2 does not have such a good spread of solutions when optimizing the GAP system. In [3] we compared NSGA-II, SPEA2, SMPSO and OMOPSO [24] on the UniMap simulator [3]. We showed that NSGA-II and SPEA2 obtain better results, in terms of quality, than the other two algorithms. It starts from a random population called the parent population. From this parent population an offspring population is generated by means of mutation and crossover (see below). The two populations (offspring and parent) are merged into a single one and the best individuals (architectural configurations) are selected according to their fitness value. The fitness value is computed considering the domination relationship and a density function. These individuals will form the new parent population and the process is repeated.

NSGA-II is not a distributed algorithm by default. To accelerate the DSE process we have changed the algorithm and now the individuals are evaluated in parallel. This is possible because the values of the objectives of an individual are required only after all the individuals are evaluated. So an entire population can be evaluated in parallel and a single synchronization point has to be established at the end of a generation. We configured the NSGA-II algorithm as follows:

Stop condition: we will observe the hypervolume progress. If there is no progress for at least *X* generations we consider that the algorithm has converged. To measure the progress we will use the following formula:

$$\text{Progress} = \sum_{i=1}^{X} (H_k - H_{k-i}) \qquad (7)$$

where H_k is the hypervolume of the current generation *k*, $X \leq k$. When this sum is smaller than a specified threshold θ the algorithm is stopped.

Population size: 100 – as recommended in [6].

Mutation: bit flip mutation with a mutation probability of $1/nparam$ [6] (where *nparam* is the number of varied parameters). In our situation the mutation probability is set to 0.05 (19 parameters).

Crossover: single point crossover, probability of crossover set to 0.9 (as specified in [6]).

Selection operator: binary tournament selection (described in [6])

Used metrics:

– Hypervolume: in a maximization problem the hypervolume is the volume enclosed between the Pareto front approximation and the axes. In a minimization problem a point has to be selected (called hypervolume reference point). The hypervolume reference point is selected at the coordinates provided by maximum values of the objectives.

– Other metrics: number of generated individuals, comparisons between the obtained Pareto fronts approximation.

5.1. Run without prior information

First of all, we start FADSE with an initial randomly generated population, without prior information. We search for the optimal SLVP-based superscalar configurations considering the same two objectives, CPI and energy consumption, as in the previous manual design space exploration. We vary the parameters presented in Table 1 with the hope to find better configurations than our manually obtained "optimal" configurations. To avoid extremes which can generate unfeasible configurations, we used the following constraints:

UL2 > DL1 + IL1

UL2_bsize ≥ DL1_bsize

UL2_bsize ≥ IL1_bsize

Where *UL2_bsize*, *DL1_bsize* and *IL1_bsize* are the block sizes for the unified L2 cache, L1 data cache and L1 instruction cache, respectively. Additionally, we limited the cache sizes by using the following hard constraints (borders):

DL1: 16 KB - 1 MB

IL1: 16 KB - 1 MB

UL2: 1 MB - 8 MB

Unfortunately, the constraints used within the initial run does not allow FADSE to efficiently explore the borders and, therefore, the configurations were not better than those obtained manually (see Figure 2) from the energy point of view. Consequently, we relaxed the minimum cache capacities as follows:

DL1: 4 KB - 1 MB

IL1: 8 KB - 1 MB

UL2: 256 KB - 8 MB

As Figure 2 shows, with relaxed borders FADSE provides significantly better configurations than our previous manual design space exploration. With these constraints the design space is reduced to 3% of the initial space, meaning 7.7 $*10^{13}$ (77 thousands of billions) configurations. Considering both objectives, the better results are influenced by the following parameters: less DL1 sets, less IL1 sets, higher decode/issue/commit width, higher ROB size, higher IQ size, higher number of MUL/DIV and higher SLVP size. Big structures will increase the energy consumption thus we do need FADSE to find the relations between different parameters. We can observe again that SLVP helps maintaining a better CPI and energy consumption when the cache sizes are reduced.

Since the exploration with relaxed borders was superior to the initial constraints, in the next experiments we used only the relaxed borders.

5.2. Run with manually obtained “optimal” configurations

The second step in our experiment consists in starting FADSE with an initial randomly generated population but containing also our manually obtained “optimal” configurations and their vicinity (with the goal to find better ones). We selected from Figure 1 the best configuration in CPI, 2UL2_DL1, the best configuration in terms of energy consumption, 1/4UL2_1/8DL1, and other two configurations which are optimal from both CPI and energy viewpoints: 1/2UL2_1/2DL1 and 1/2UL2_1/4DL1. We also considered the vicinities of these four configurations by varying the SLVP size, L1 data cache size and L2 unified cache size one step up and down. Thus, we started FADSE again with randomly generated population but containing also our 24 selected configurations: the “optimal” manual configurations and their vicinities (some of them are overlapped). Figure 2 shows the obtained Pareto fronts after 25 generations by the first three runs (initial run, run with relaxed borders and run with initial good configurations) compared with the manually obtained configurations.

In terms of CPI all the runs find much better solutions than the manually obtained configurations. The run with relaxed borders clearly finds better configurations than the ones obtained through manual exploration and also better than the ones found during the initial run (restrictive constraints). The obtained solutions are distributed evenly along the Pareto approximated front.

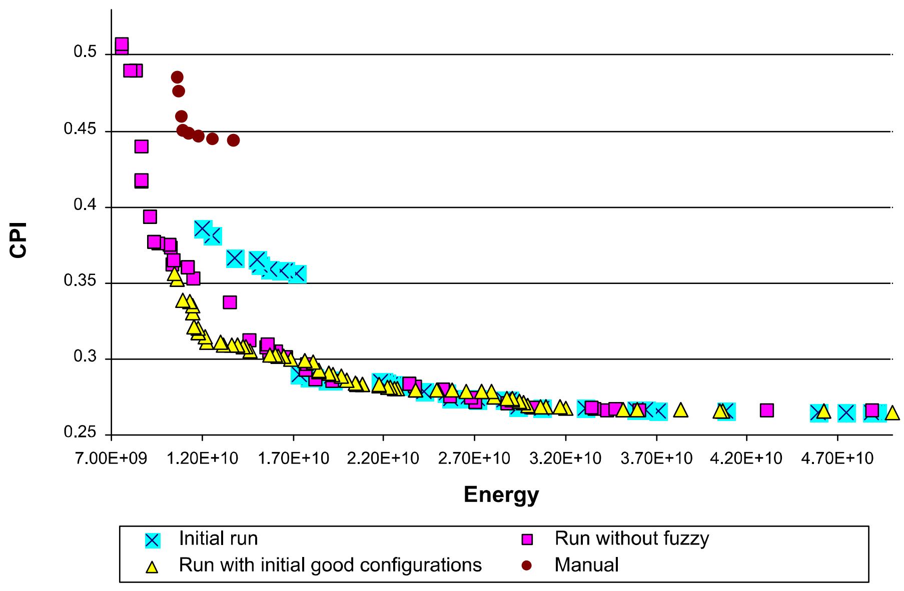

Figure 2. Pareto fronts comparison

Inserting good configurations into the initial population provides also good results but it is not able to explore the area with very low energies. It obtains better results than the run with relaxed borders in the vicinity of energy 1.20E+10 [$W \cdot cycles$]. Low energy configurations are not found probably because the initial configurations were better (and we have observed this on our analysis of the Pareto front approximation evolution over the generations) than all the other individuals inserted randomly in the population. So all these good individuals survived until the next generation and most of the offspring were generated from them thus loosing diversity. The mutation operator with a probability of 0.05 of changing one parameter has a small chance of influencing significantly the produced offspring, leading to a reduction of diversity.

5.3. Run with knowledge expressed through fuzzy rules

We are using fuzzy rules to allow the designer to express knowledge. The information provided by these fuzzy rules is then used during the search process to guide the DSE algorithm. For this purpose we have included the jFuzzyLogic library (http://jfuzzylogic.sourceforge.net) in FADSE. A user can define rules in a standard FCL file (IEC 61131 part 7). In this article we have developed and implemented the following rules derived from our experience in computer architecture design:

IF Number_Of_Physical_Register_Sets IS *small/big* THEN Decode/Issue/Commit_Width IS *small/big*

IF SLVP_size IS *small/big* THEN L1_Data_Cache IS *big/small*

We have selected the Mamdani-type fuzzy systems [17]. These imply the following steps that need to be carried to extract information: fuzzification of the input variables, evaluating the rules, aggregating the outputs and then defuzzification. The fuzzification was done using trapezoidal functions. For the evaluation the *min* function was used for “and” and *max* for “or”. For inference the Mamdani implication has used (*min*). The rule aggregation was performed using the Mamdani aggregation (*max*). This system was selected because of its popularity.

Two different mutation operators were used. Both of them are based on the bit flip mutation. To preserve diversity, the information provided by the fuzzy rules is not always taken into consideration. To obtain this, a probability of applying the fuzzy information (called fuzzy probability) is used. The only difference between the implemented methods is how the probability to apply the information provided by the fuzzy rules is computed.

In the simple implementation this fuzzy probability is constant during the run of the algorithm and it is set to be equal with the probability of mutation (*mutation_prob*). If the fuzzy rule is not applied the algorithm switches to the classical bit flip mutation for the current parameter. The second implementation uses a Gaussian probability so there is a higher chance to apply the fuzzy rules for the first generations. As the DSE process runs, the probability to apply the fuzzy rules decreases to a value close to *mutation_prob*. We have selected the parameters of the Gaussian function such that at generation 5 the function is close to 0. The Gaussian function is then translated so that the minimum is close to *mutation_prob*. The final form of the function is shown below:

$$f(x)_{final} = (1 - mutation\_prob) \cdot e^{\frac{-(x)^2}{2 \cdot (150)^2}} + mutation\_prob \qquad (8)$$

where x increases with one for each individual generated by the algorithm. The result of this function is further multiplied by 0.8 and by the membership value [27] to obtain a maximum value less than 1.

Figure 3 presents the results obtained with fuzzy information compared with the previous results.

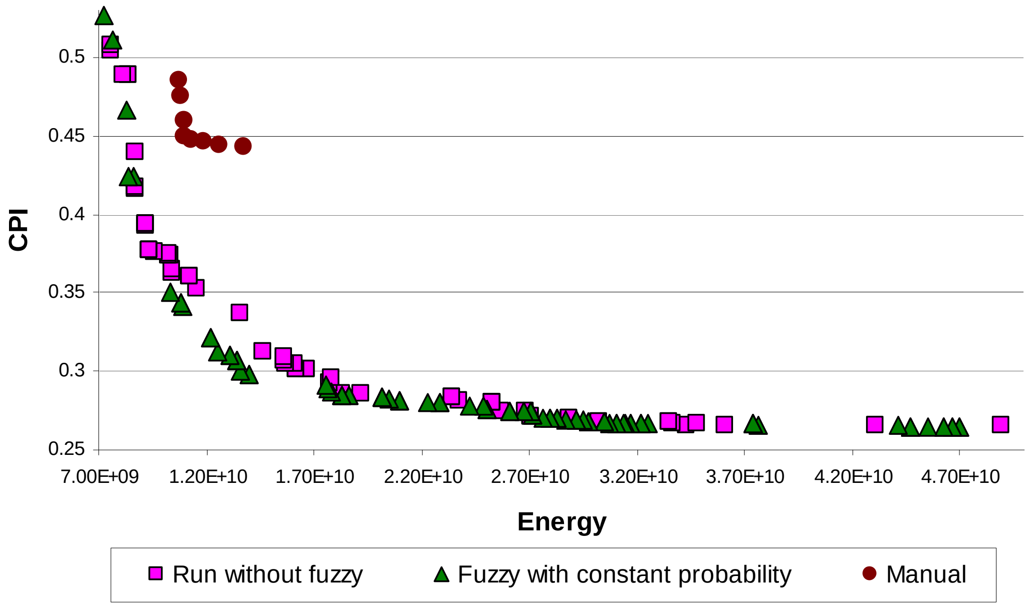

Figure 3. Pareto front comparisons between the run with fuzzy rules and the run with relaxed borders

We have selected the run with relaxed borders (called "run without fuzzy") as the reference run since it has found solutions all along the approximated Pareto front. In Figure 3 we compare the run with fuzzy information (constant probability to apply the fuzzy rules) with the run without fuzzy. It can be easily observed that the run with fuzzy information obtains very good results. Figure 3 also shows that the run with fuzzy rules finds better results in the vicinity of energy 1.20E+10 [$W \cdot cycles$] than the run without fuzzy information. We have also compared the run with fuzzy rules with the one with initial good configurations and we observed that in fact the later obtains a few individuals which are slightly better.

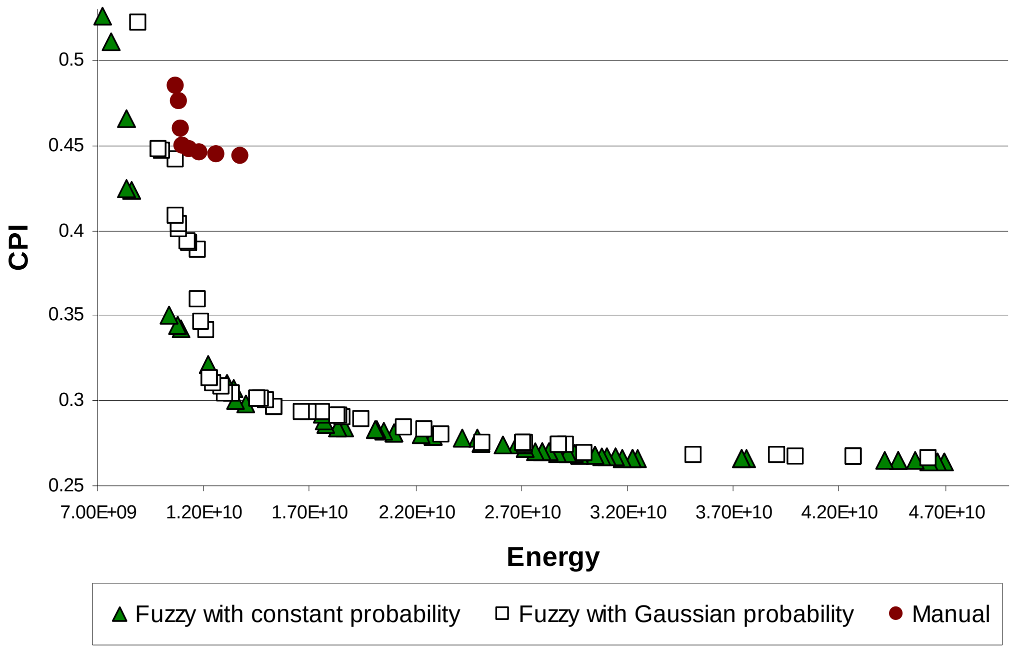

Figure 4. Pareto fronts comparison between the runs with fuzzy rules

Figure 4 performs a comparison between the results obtained with fuzzy information but with different methods of calculating the probability to use the information provided by them. The run with constant probability finds better individuals in the area with low energy. Having an almost 80% probability to apply the rules during the first generations might lead to a loss in diversity of the individuals on the parameters influenced by the rules. This fact might explain the poorer results.

Figure 5 gives us two types of information: about the convergence of the algorithms and about the quality of results. It can be observed that the algorithms tend to stop the rapid evolution after 15 generations (initial run is an exception). The algorithms were run until generation 25 due to time constraints.

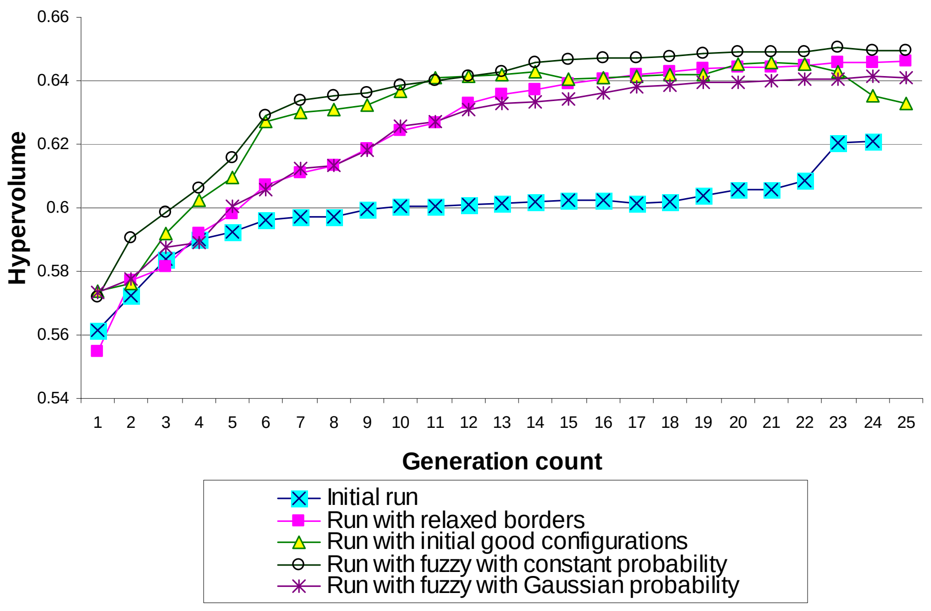

Figure 5. Hypervolume comparison

After a comparison of the hypervolume values we can conclude that: the initial run with restricted borders obtained the worse results. Relaxing the borders considerably improved the quality of results even though the size of the design space has become larger by a factor of around 2 (from $3.8*10^{13}$ to $7.7*10^{13}$). Having good configurations inserted in the initial population can lead to very good results but it starts to perform worse after generation 14, rising a bit after generation 19 and then dropping dramatically after generation 23. Observing the obtained Pareto front approximation and its evolution we have concluded that the algorithm tends to focus on a smaller area of the space, falling into local minima. This can be explained by the lack of diversity of the initial configurations, since all the individuals inserted differ on only two parameters from a total of 19.

The run with a constant probability of accepting the results from the fuzzy rules provided the best results. The run with a Gaussian probability of applying the information provided by the fuzzy rules had a similar behavior at the beginning with the run using relaxed borders. After generation 12 the results are slightly worse. We can conclude that imposing a high probability of the rules will reduce diversity, especially with a small number of rules. In our previous work more rules were used and the membership functions had many intervals (associated linguistic terms) [13]. In this situation the runs with Gaussian probability provided better results.

It can be observed that using some extra knowledge (initial configurations or fuzzy rules) makes the algorithm start from a better initial population (see the hypervolume values at generation 1) and, as a consequence, the algorithm's convergence speed is better.

The hypervolume corresponding to the run with a constant probability of applying the fuzzy rules at generation 15, is reached by the run with relaxed borders only at generation 24. This is a great improvement. In our experiments, running one generation on 96 cores belonging to an Intel Xeon powered HPC system, with cores running at 2GHz, takes around one day. Running with fuzzy rules we achieved the same results 9 days earlier (36% faster) than without fuzzy rules. Additionally, after the same amount of time (25 generations) the hypervolume reached by the run with fuzzy is never reached by the simple run. If more qualitative information would have been provided through the fuzzy rules we do expect even bigger improvements. A single experiment takes around 25 days. Running all the five experiments took over 4 months of simulation on a HPC system using 96 cores.

All the runs evaluate roughly the same amount of individuals: around 2200 from the 2500 individuals sent for evaluation; the rest are reused from the database (12% reuse degree). This means that the produced offspring are almost all of them new/different individuals. This behavior is caused by the extremely large design space. In previous explorations on different simulators (smaller design space – 10^6) around 60% reuse degree was observed [2, 12].

After analyzing the best results obtained during the automatic DSE process we extracted the parameter values for optimal configurations from either high performance or low energy viewpoints. The results are presented in Table 2. As it can be observed, some parameters like Decode width and SLVP size must have high values in order to obtain both high performance and low energy.

Parameter		**For high performance**	**For low energy**
DL1 cache	Sets	2048	32
	Block size (bytes)	256	64
	Associativity	2	2
IL1 cache	Sets	1024	32
	Block size (bytes)	16	256
	Associativity	8	1
UL2 cache	Sets	8192	256
	Block size (bytes)	256	256
	Associativity	4	16
SLVP (entries)		4096-8192	4096-8192
Decode / Issue / Commit width		32 / 16 / 32	16-32 / 4-8 / 16-32
ROB / LSQ / IQ size (entries)		1024 / 512-1024 / 128	256 / 32-64 / 64
Number of physical register sets (int / fp)		8 / 8	2 / 2
Int / fp ALU		8 / 8	8 / 2
Int / fp MUL/DIV		8 / 8	8 / 8

Table 2. Parameter values for optimal configurations

6. Conclusions and Further Work

We have observed that the SLVP helps maintaining a better CPI and energy consumption when the cache sizes are reduced. Therefore, the optimal configurations, obtained by both the manual and automatic design space exploration of our SLVP-based superscalar architecture, have lower cache sizes than the baseline architecture without SLVP.

FADSE is able to find good configurations by evaluating a very small percentage of the total search space. We reduced the number of evaluated configurations to only 2500, representing $3*10^{-11}$% of the huge constrained design space of 77 thousands of billions configurations. The experimental results show that our automatic design space exploration provides significantly better configurations than our previous manual design space exploration.

Starting FADSE with initial good configurations can accelerate the DSE process. It is recommended that the configurations differ on multiple parameters so that diversity is preserved. In our situation the configurations differed only on three parameters thus leading to a loss of diversity and finally it could not explore the entire Pareto front.

Using fuzzy rules can considerably accelerate the DSE process (9 days earlier to reach the same result in our situation). Also the obtained results are better than the ones obtained with no prior information after the same amount of time. In this concrete optimization process the constant probability to apply the fuzzy rules lead to better results. In our previous work – where the number of rules was higher and had more linguistic terms associated to the membership functions – we have obtained better results by modulating the probability with a Gaussian function during the generations. With a larger number of rules the individuals are mutated into a more diverse population. Thus, forcing the rules to be applied often does not lead to a loss of diversity in the population.

We plan to repeat these experiments on SLVP-based SMT and multi-core architectures. Other further work possibilities are to access the SLVP only in the case of miss in both the L1 and L2 data caches, to index the SLVP table with the memory address instead of the instruction address, to exploit an N-value locality instead of 1-value locality as we are currently exploiting, to evaluate set-associative SLVP configurations and yet another one to design and implement an adaptive dynamic run-time thermal manager (temporarily deactivating the SLVP unit, voltage scaling, frequency scaling, migrating computation, etc.). We also plan to explore, as the third objective, the die size needed by the memory architecture.

Acknowledgments

This work was partially supported by the POSDRU financing contract 7706 and by the Romanian National Council of Academic Research through the grant CNCSIS no. 485/2009-2011. Also we would like to thank Camil Bancioiu for his support with the M-SIM connector and to Ciprian Radu for his useful comments.

References

[1] Brooks, D., Tiwari, V., and Martonosi, M.: 'Wattch: A Framework for Architectural-Level Power Analysis and Optimizations', Proc. of the 27th Int. Symposium on Computer Architecture, Vancouver, Canada, June 2000, pp. 83-94.

[2] Calborean, H., Jahr, R., Ungerer, T., and Vintan, L.: 'Optimizing a Superscalar System using Multi-objective Design Space Exploration', Proc. of the 18th Int. Conf. on Control Systems and Computer Science (CSCS), Bucharest, Romania, May 2011, pp. 339-346.

[3] Calborean, H.: 'Multi-Objective Optimization of Advanced Computer Architectures using Domain-Knowledge', PhD Thesis, "Lucian Blaga" University of Sibiu, Romania, 2011 (PhD Supervisor: Prof. Lucian Vintan, PhD), Sibiu, Romania, 2011.

[4] Chang, S.C., Li, W.Y.H., Kuo, Y.J., and Chung, C.P.: 'Early Load: Hiding Load Latency in Deep Pipeline Processor', Proc. of the Asia-Pacific Computer Systems Architecture Conference (ACSAC), Taiwan, August 2008, pp. 1-8.

[5] Chen, J., Dubois, M., and Stenström, P.: 'Integrating Complete-System and User-Level Performance/Power Simulators: the SimWattch Approach', Proc. of the IEEE Int. Symposium on Performance Analysis of Systems and Software, Austin, Texas, USA, March 2003, pp. 1-10.

[6] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T.: 'A Fast and Elitist Multiobjective Genetic algorithm: NSGA-II', IEEE Transactions On Evolutionary Computation, VOL. 6, 2002, NO. 2, pp. 182-197.

[7] Desmet, V., Girbal, S., Temam, O., and France, B. F.: 'Archexplorer. org: Joint compiler/hardware exploration for fair comparison of architectures', Proc. of the 6th HiPEAC Industrial Workshop, Paris, France, November 2008.

[8] Durillo, J. J., Nebro, A. J., Luna, F., Dorronsoro, B., and Alba, E.: 'jMetal: A Java Framework for Developing Multi-Objective Optimization Metaheuristics', E.T.S.I. Informatica, Campus de Teatinos: Departamento de Lenguajes y Ciencias de la Computacion, University of Malaga, 2006.

[9] Gellert, A., Florea, A., and Vintan, L.: 'Exploiting Selective Instruction Reuse and Value Prediction in a Superscalar Architecture', Journal of Systems Architecture, Elsevier, Volume 55, 2009, Issue 3, pp. 188-195.

[10] Gellert, A., Palermo, G., Zaccaria, V., Florea, A., Vintan, L., and Silvano, C.: 'Energy-Performance Design Space Exploration in SMT Architectures Exploiting Selective Load Value Predictions', Int. Conf. on Design, Automation and Test in Europe (DATE 2010), Dresden, Germany, March 2010, pp. 271-274.

[11] Hilton, A. D.: 'Energy Efficient Load Latency Tolerance: Single-Thread Performance For The Multi-Core Era', 2010, Publicly accessible Penn Dissertations. Paper 188, http://repository.upenn.edu/edissertations/188.

[12] Jahr, R., Ungerer, T., Calborean, H., and Vintan, L.: ‘Automatic Multi-Objective Optimization of Parameters for Hardware and Code Optimizations’, Proc. of the Int. Conf. on High Performance Computing & Simulation, Istanbul, Turkey, July 2011, pp. 308-316.

[13] Jahr, R., Calborean, H., Ungerer, T., and Vintan, L.: ‘Boosting Design Space Explorations with Existing or Automatically Learned Knowledge’, accepted for publication at 16th International GI/ITG Conference on “Measurement, Modelling and Evaluation of Computing Systems” and “Dependability and Fault-Tolerance” (MMB & DFT 2012), Kaiserslautern, Germany, March 19 - 21, 2012.

[14] Jia, Z. J., Pimentel, A. D., Thompson, M., Bautista, T., and Núnez, A.: ‘NASA: A generic infrastructure for system-level MP-SoC design space exploration’, in Embedded Systems for Real-Time Multimedia (ESTIMedia), 2010 8th IEEE Workshop on, 2010, pp. 41–50.

[15] Kang, S. and Kumar, R.: ‘Magellan: a search and machine learning-based framework for fast multi-core design space exploration and optimization’, in Proceedings of the conference on Design, automation and test in Europe, Munich, Germany, 2008, pp. 1432-1437.

[16] Lipasti, M. H., Wilkerson, C. B., and Shen, J. P.: ‘Value Locality and Load Value Prediction’, Proc. of the 7th Int. Conf. on Architectural Support for Programming Languages and Operating Systems, Cambridge, Massachusetts, USA, October 1996, pp. 138-147.

[17] Mamdani, E. H., and Assilian, S.: ‘An experiment in linguistic synthesis with a fuzzy logic controller’, International Journal of Man-Machine Studies, 7(1), 1975, pp.1–15.

[18] Mariani, G., Palermo, G., Silvano, C., and Zaccaria, V.: ‘Meta-model Assisted Optimization for Design Space Exploration of Multi-Processor Systems-on-Chip’, Proc. of the 12th Int. Euromicro Conf. on Digital System Design, Architectures, Methods and Tools. DSD ’09, IEEE Computer Society, Patras, Greece, August 2009, pp. 383–389.

[19] Mutlu, O., Kim, H., and Patt, Y. N.: ‘Address-Value Delta (AVD) Prediction: A Hardware Technique for Efficiently Parallelizing Dependent Cache Misses’, IEEE Transactions on Computers, Vol. 55, No. 12, 2006, pp. 1491-1508.

[20] Nebro, A., Durillo, J., Garcia-Nieto, J., Coello, C. A., Luna, F., and Alba, E.: ‘SMPSO: A new PSO-based metaheuristic for multi-objective optimization’, in Proceedings of the IEEE Symposium Series on Computational Intelligence, 2009, pp. 66–73.

[21] Sarangi, S. R., Liu, W., Torrellas, J., and Zhou, Y.: ‘ReSlice: Selective Re-Execution of Long-Retired Misspeculated Instructions Using Forward Slicing’, Proceedings of the 38th annual IEEE/ACM International Symposium on Microarchitecture, Barcelona, Spain, November 12-16, 2005, pp.257-270.

[22] Sharkey, J., Ponomarev, D., and Ghose, K.: ‘M-SIM: A Flexible, Multithreaded Architectural Simulation Environment’, Technical Report CS-TR-05-DP01, Department of Computer Science, State University of New York at Binghamton, October 2005.

[23] Shivakumar, P., and Jouppi, N. P.: ‘Cacti 3.0: An Integrated Timing, Power, and Area Model’, WRL Research Report, Aug 2001, USA.

[24] Sierra, M. R., and Coello, C. A. C.: ‘Improving PSO-based multi-objective optimization using crowding, mutation and e-dominance’, Evolutionary Multi-Criterion Optimization, Springer-Verlag, 2005, pp. 505-519.

[25] Silvano, C., Fornaciari, W., Palermo, G., Zaccaria, V., Castro, F., Martinez, M., Bocchio, S., Zafalon, R., Avasare, P., Vanmeerbeeck, G., Ykman-Couvreur, C., Wouters, M., Kavka, C., Onesti, L., Turco, A., Bondi, U., Mariani, G., Posadas, H., Villar, E., Wu, C., Dongrui, F., Hao, Z., and Shibin, T.: ‘MULTICUBE: Multi-Objective Design Space Exploration of Multi-Core Architectures’, Proc. of the IEEE Int. Annual Symposium on VLSI, Lixouri Kefalonia, Greece 2010, pp. 488–493.

[26] Vintan, L., Florea, A., and Gellert, A.: ‘Focalising dynamic value prediction to CPU’s context’, IEE Proceedings – Computers and Digital Techniques, Stevenage, United Kingdom, Volume 152, Issue 4 (July), 2005, pp. 457-536.

[27] Zadeh, L. A.: ‘Fuzzy Sets’, Journal of Information and Control, Vol. 8, No. 3, 1965, pp. 338-353.

[28] Zitzler, E., Laumanns, M., and Thiele, L.: ‘SPEA2: Improving the strength Pareto evolutionary algorithm’, in Eurogen, 2001, pp. 95–100.