A System-Level Methodology for Fast Multi-Objective Design Space Exploration

G. Palermo C. Silvano S. Valsecchi V. Zaccaria

Politecnico di Milano, Dipartimento di Elettronica e Informazione, 20133 Milano, Italy

ABSTRACT

In this paper, we address the problem of the efficient exploration of the architectural design space for parameterized systems. Since the design space is multi-objective, our aim is to find all the Pareto-optimal configurations that represent the best design trade-offs by varying the architectural parameters of the target system. In particular, the paper proposes a Design Space Exploration (DSE) framework based on a random search algorithm that has been tuned to efficiently derive Pareto-optimal curves. The reported design space exploration results have shown a reduction of the simulation time of up to two orders of magnitude with respect to full search strategy, while maintaining an average accuracy within 3%.

Categories and Subject Descriptors

C.3 [**Special-Purpose and Application-Based Systems**]: Embedded Systems

General Terms

Design, Performance

Keywords

Design Space Exploration, Low-Power Design, Embedded Systems, System-Level Methodologies

1. INTRODUCTION

Decreasing energy consumption without a relevant impact on performance is a 'must' during the design of a broad range of embedded applications. Evaluation of energy-delay metrics at the system-level is of fundamental importance for embedded applications characterized by low-power and high-performance requirements.

GLSVLSI'03, April 28–29, 2003, Washington, DC, USA.

Given the application-specific functionality, the design of an embedded system requires the definition of the best architecture mainly in terms of core processor, degree of Instruction Level Parallelism (ILP), number of levels in the memory hierarchy, cache-related parameters, system-level bus topology, width of address and data busses, etc.. To achieve this goal, an approach based on the full search of the optimal architectural parameters at the system-level with respect to the energy-delay cost function can be computationally very costly due to the long simulation time required to explore the wide space of parameters.

In general, parameterized embedded System-On-Chip architectures must be optimally tuned to find the best energy-delay trade-offs for the given classes of applications. The value assignment to each one of the system-level parameters can significantly impact the overall performance and power consumption of the given embedded architecture. The problem addressed in this paper consists of defining a Design Space Exploration (DSE) framework to efficiently explore the multi-objective design space in order to find a good approximation of Pareto-optimal configurations representing the best compromise between the interesting design objectives, mainly energy and delay. The proposed exploration technique is based on the application of a random search algorithm (namely, Random Search Pareto - RSP), that has been tuned to efficiently derive a good approximation of Pareto-optimal curves.

The analysis of the proposed algorithm has been carried out for a number of multimedia applications simulated on our target parameterized embedded architecture. The results derived from the proposed algorithm have been compared to the results derived from a full search algorithm. The reported exploration results have shown a reduction of up to two orders of magnitude with respect to the full search, while maintaining a good level of accuracy (within 3% on average).

The rest of the paper is organized as follows. A review of the most significant works appeared in literature concerning the DSE problem is reported in Section 2. The DSE problem is stated in Section 3 along with the problem of the approximation of Parte curves. The proposed DSE framework is described in Section 4, while Section 5 discusses the experimental results carried out to evaluate the efficiency of the proposed framework. Finally some concluding remarks have been reported in Section 6.

2. BACKGROUND

Several system-level estimation and exploration methods have been recently proposed in literature targeting power-performance tradeoffs from the system-level standpoint.

The SimplePower approach [1] can be considered one of the first efforts to evaluate the different contributions to the energy budget at the system-level. The *Avalanche* framework presented in [2] evaluates simultaneously the energy-performance tradeoffs for software, memory and hardware for embedded systems. The work in [3] proposes a system-level technique to find low-power high-performance superscalar processors tailored to specific user applications. More recently, the Wattch architectural-level framework has been proposed in [4] to analyze power vs. performance tradeoffs with a good level of accuracy with respect to lower-level estimation approaches. Low-power design optimization techniques for high-performance processors have been investigated in [5] from the architectural and compiler standpoints. A trade-off analysis of power performance effects of SOC (System-On-Chip) architectures has been recently presented in [6], where the authors propose a simulation-based approach to configure the parameters related to the caches and the buses.

3. DESIGN SPACE EXPLORATION

Among the factors influencing the success of an electronic product, time-to-market is becoming the most crucial one. To meet such time constraints, modern design techniques are oriented towards the use of customizable System-On-Chip (SOC) platforms [7] where a stable microprocessor-based architecture can be easily extended and customized for a range of applications, and delivered to customers for quick deployment and, possibly, at low cost.

Even considering a simple embedded microprocessor-based architecture composed of the CPU and the memory hierarchy, the identification of the optimal system configuration by trading off power and performance still leads to the analysis of too many alternatives. The overall goal of this work aims at overcoming such problems by providing a methodology and a design framework to drive the designer towards optimal solutions in a cost-effective manner.

3.1 Parameterized Design Space Definition

We define the *design space* as the set of all the feasible architectural implementations of a platform. Associated with each point of the design space, there are a set of evaluation functions (or *metrics*). The *design evaluation space* is the multi-dimensional space spanned by these evaluation functions. For the embedded system category analyzed in this work, a reasonable set of parameters affecting the performance and the energy consumption is composed of the number and configuration of the levels of the memory hierarchy, the address busses, etc.

The problem afforded in this paper consists of platform optimization, that is searching for the best design, i.e., an implementation that optimizes all the objectives within the design evaluation space. However, the optimization problem involves the minimization (maximization) of *multiple objectives* making the definition of optimality not unique.

To address this problem, let us introduce the definition of *Pareto point* for a minimization problem [8]. A Pareto point is a point of the design space, for which there is no other point with at least an inferior objective, all others being inferior or equal. Obviously, a Pareto point is a global optimum in the case of a monodimensional design space, while in the case of multi-dimensional design evaluation space, the Pareto points form a trade-off curve or surface called Pareto curve.

In general, Pareto points are solutions to *multi-objective* or *constrained* optimization problems. For example, we can be interested in minimizing the power consumption under a delay constraint or viceversa. The solution of this problem is straightforward if the Pareto curve is available. However, a Pareto curve for a specific platform is available only when all the points in the design space have been characterized in terms of objective functions. This is often unfeasible due to the cardinality of the design space and to the long simulation time needed for computing the evaluation functions.

Our target problem consists of finding a good approximation of the Pareto curves by trading off the accuracy of the approximations and the time needed for their construction.

3.2 Approximation of Pareto Curves

The most trivial approach to determine the Pareto-optimal configurations into a large design space with respect to a multi-objective design optimization criteria consists of the comprehensive exploration of the configuration space. This *brute force* approach can be feasible only if the number of parameters in the configuration space is very limited. On the contrary, it is quite common to find a design space composed of tens of parameters, leading to an exponential analysis time. Thus, traditional heuristics must be used.

Heuristics can be divided in two main categories: *local search algorithms* and *global search algorithms*. Local search algorithms are typically used for finding local solutions to an optimization problem with a high convergence speed. These algorithms are usually characterized by an iterative process in which each new point to be analyzed is in the neighborhood of the points previously analyzed.

Global search algorithms are typically used for finding global solutions to optimization problems; they are fundamentally based on rules that allow escaping from local minima by means of random point selection (e.g., Random Search [9] or Genetic Algorithms).

Very few approaches have been introduced in the literature for approximate Pareto-curve curve construction of computer architecture design [10] [11].

Platune [10] is an optimization framework that exploits the concept of parameter independence to individuate approximate Pareto curves without performing the exhaustive search over the whole design space. The main drawback of this approach is that parameter independence must be specified by the user by means of a dependency graph since no automatic method is proposed for such task. Platune is not a modular framework since it allows only the exploration of a MIPS based system and its goodness has not been compared with simpler approaches such as random search or full parameter space exploration.

Palesi et al. [11] extended Platune with genetic algorithms to derive approximation of Pareto curves. However, their approach is always based on an a-priori parameter dependency graph to be given by the user and it is compared only with the default Platune policy.

In this study, we introduce and analyze a framework based on the Random Search Pareto (RSP) algorithm to compute a good approximation of Pareto curves in an efficient manner

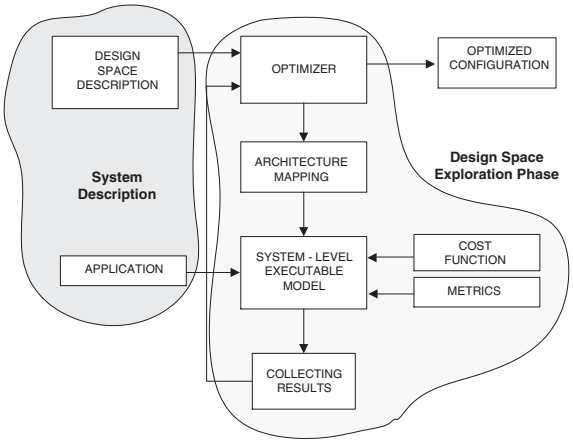

Figure 1: The proposed design space exploration framework

without an a-priori knowledge of the system. The results of the proposed algorithm have been compared to the results derived from the Full Search (FS) algorithm. In our best knowledge, our work represents the first attempt in the literature to use a random search algorithm to generate Pareto curves for the multi-objective design space exploration problem.

4. PROPOSED DESIGN SPACE EXPLORATION FRAMEWORK

The overall goal of this work aims at providing a methodology and a retargetable tool to drive the designer towards near-optimal solutions, with the given multiple constraints, in a cost-effective fashion. The final product of the framework is a Pareto curve of configurations within the design evaluation space of the given application.

The proposed framework is shown in Figure 1. The framework receives as input a description of the possible design configurations (i.e. the target *design space*) and the application for which the optimal configuration must be found with respect to different metrics. The target design space and the object functions (metrics) are strictly paired with the target architectural simulator used to estimate the design evaluation space.

The approximation of the Pareto curve is performed by a Random Search Pareto (RSP) algorithm derived from *Monte Carlo* methods. In general, the main characteristic of *Monte Carlo* methods is the use of random sampling techniques to come up with a solution of the target problem. The random sampling technique has been proved to be one of the best techniques to avoid falling into local minima. The RSP algorithm receives as input the number of maximum simulations N_{MAX} to be generated. The design space is thus sampled by the algorithm and an approximate Pareto curve is constructed by filtering the Pareto points of the N_{MAX} configurations within the design evaluation space.

In the next section, the validation of the performance and the accuracy of the RSP algorithm in generating Pareto curves has been evaluated in the following way. The performance has been evaluated in terms of number of simulations, while the accuracy has been computed as the error between the estimated Pareto curve (Random Search Pareto, RSP) and the actual Pareto curve (Full Search Pareto, FSP) over a large set of benchmarks. Furthermore, in the next section, we will show a case study of the application of our methodology to the constrained optimization of a given application. In this case, we consider the error of approximated pareto curves defined in [12].

5. EXPERIMENTAL RESULTS

In this section, we present the experimental results obtained by applying the proposed DSE framework based on the RSP algorithm to optimize a superscalar microprocessor-based system.

5.1 Target System Architecture

In general, a superscalar architecture is composed of many parameters, so that the design space to explore is quite large. Our analysis has been focused on those design parameters significantly impacting the performance, the energy consumption and the silicon area occupation. Each instance of the virtual architecture has been described in terms of size, block size and associativity of I/D L1 caches and unified L2 cache, number of integer and floating point ALUs and multipliers, and finally issue width size. Wattch simulator [4] has benn used as our target architectural simulator, which has been integrated with an area model developed by our research group [12]. Wattch integrates also cache energy and delay models.

5.2 Validation of the methodology

In this subsection, we report the results in terms of the efficiency and accuracy of the application of our DSE methodology to a selected set of benchmarks. The chosen set of benchmarks is composed by a set of DCT transforms and FIR filters as well as other numerical algorithms written in C language. The validation flow, which has been used, is composed of two parallel sub-flows:

1. Full Search Flow (*FSF*): This flow derives the Pareto curves for the benchmarks by using the Full Search algorithm.

2. Random Search Flow (*RSF*): This flow derives the Pareto curves for the benchmarks by using the RSP algorithm.

Each benchmark has been optimized independently with the two flows and the resulting Pareto curves have been compared. In the Full Search case, the optimizer analyzes the entire design space, so the number of simulations to be executed is 196608. This corresponds to approximately 370 hours of simulations for the entire set of benchmarks. For what concerns the RSP algorithm, it has been validated by varying the number of maximum simulations N_{MAX} from 10 to 100.000 in a logarithmic fashion. Figure 2 shows the behavior of the minimum, average and maximum error of the approximated RSP curves with respect to the full search Pareto curves. As can be seen, for $N_{MAX} = 100$, the average error falls under 3%, for $N_{MAX} = 1000$ under 7%, and for $N_{MAX} = 10000$ under 1%. While the curves describing the minimum, average and maximum errors show a decreasing trend, they presents some oscillations due to the random nature of the search. RSP Simulation time is up to two orders of magnitude faster than FSP, as measured on a Linux Workstation provided with the Intel Pentium 4 at

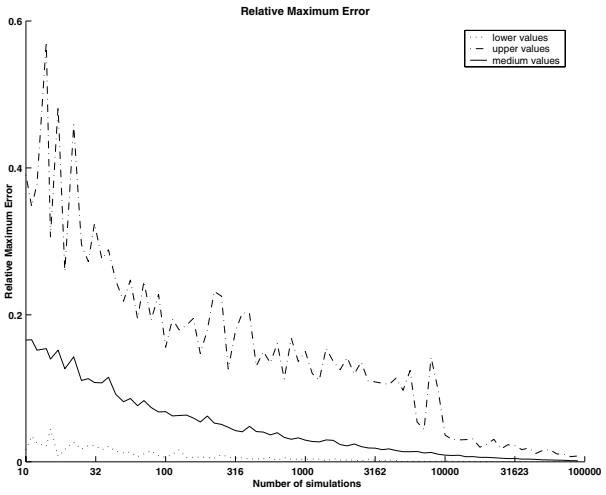

Figure 2: The minimum, average and maximum error in the approximation of the pareto curves

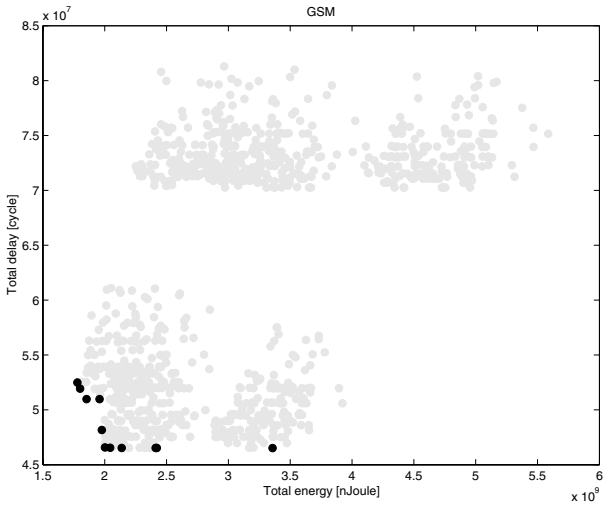

Figure 3: Exploration results of the energy-delay design evaluation space

1.7 GHz clock speed. RSP Simulation time is up to two orders of magnitude faster than FSP, as measured on a Linux Workstation provided with the Intel Pentium 4 at 1.7 GHz clock speed.

5.3 Case study

This section presents the results obtained by applying the RSP algorithm to the optimization of the *GSM* encoding benchmark. Our DSE problem is to find the optimal configuration in terms of energy with a constraint on the delay of 50 millions of cycles for the encoding of a 73760 byte of data. To solve this problem, we compute the RSP curve of the benchmark for the Energy and Delay design evaluation space by applying the proposed algorithms with $N_{MAX} = 1000$. Figure 3 shows the scatter plot of all the points generated by the RSP algorithm (light gray) and the corresponding Pareto points (dark gray). Among the Pareto points, we have found the optimal RSP point characterized by the following metric values: Energy = 1.97 [J], Delay = 48x10^6 [cycles], Area = 86 [mm^2]. This point fulfills the constraints imposed by the problem in terms of delay and it minimizes the energy. The analysis on the accuracy of the algorithms performed in the previous section suggests that this point is very close to the Full Search corresponding point. The RSP analysis required a simulation time of three days, while the FS analysis would have required 516 days of simulation.

6. CONCLUSIONS

In this paper, a DSE framework has been proposed to efficiently derive Pareto-optimal curves. The framework is based on a random search exploration technique, that has been compared to the full search exploration. For the given set of benchmarks, the RSP techniques is up to two orders of magnitude faster than the full search, while maintaining its accuracy within 3% on average. Future developments of our works aim at evaluating in our DSE framework other heuristic search algorithms, such as the Tabu Search and Simulated Annealing algorithms.

7. REFERENCES

[1] N. Vijaykrishnan, M. Kandemir, M.J. Irwin, H.S. Kim, and W. Ye. Energy-driven integrated hardware-software optimizations using simplepower. In *ISCA 2000: 2000 International Symposium on Computer Architecture*, Vancouver BC, Canada, June 2000.

[2] Y. Li and J. Henkel. A framework for estimating and minimizing energy dissipation of embedded hw/sw systems. In *DAC-35: ACM/IEEE Design Automation Conference*, June 1998.

[3] T. M. Conte, K. N. Menezes, S. W. Sathaye, and M. C. Toburen. System-level power consumption modeling and tradeoff analysis techniques for superscalar processor design. *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 8(2):129–137, Apr. 2000.

[4] David Brooks, Vivek Tiwari, and Margaret Martonosi. Wattch: a framework for architectural-level power analysis and optimizations. In *Proceedings ISCA 2000*, pages 83–94, 2000.

[5] N. Bellas, I. N. Hajj, D. Polychronopoulos, and G. Stamoulis. Architectural and compiler techniques for energy reduction in high-performance microprocessors. *IEEE Transactions on Very Large Scale of Integration (VLSI) Systems*, 8(3), June 2000.

[6] Tony D. Givargis, Frank Vahid, and Jörg Henkel. Evaluating power consumption of parameterized cache and bus architectures in system-on-a-chip designs. *IEEE Transactions on Very Large Scale of Integration (VLSI) Systems*, 9(4), August 2001.

[7] K. Keutzer, S. Malik, A. R. Newton, J. Rabaey, and A. Sangiovanni-Vincentelli. System level design: Orthogonolization of concerns and platform-based design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 19(12):1523–1543, December 2000.

[8] A. Aho, J. Hopcroft, and J. Ullman. *Data Structures and Algorithms*. Addison-Wesley, Reading, MA, USA, 1983.

[9] Anatoly A. Zhigljavsky. *Theory of global random search*, volume 65. Kluwer Academic Publishers Group, Dordrecht, 1991.

[10] Tony D. Givargis and Frank Vahid. Platune: a tuning framework for system-on-a-chip platforms. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 21(11):1317–1327, November 2002.

[11] M. Palesi and T Givargis. Multi-objective design space exploration using genetic algorithms. In *Proceedings of the Tenth International Symposium on Hardware/Software Codesign, 2002. CODES 2002*, May 6–8 2002.

[12] S. Valsecchi. *An algorithm for the efficient exploration of the architectural design space for microprocessor-based systems*. M.S. Thesis, Dipartimento di Elettronica e Informazione, Politecnico di Milano, 2002.