Experimental Evidence of Power Efficiency due to Architecture in Cellular Processor Array Chips

Ricardo Carmona-Galán, Jorge Fernández-Berni, Ángel Rodríguez-Vázquez
Instituto de Microelectrónica de Sevilla (IMSE-CNM), CSIC-Universidad de Sevilla, Spain
Email: rcarmona@imse-cnm.csic.es

***Abstract*—Speeding up algorithm execution can be achieved by increasing the number of processing cores working in parallel. Of course, this speedup is limited by the degree to which the algorithm can be parallelized. Equivalently, by lowering the operating frequency of the elementary processors, the algorithm can be realized in the same amount of time but with measurable power savings. An additional result of parallelization is that using a larger number of processors results in a more efficient implementation in terms of GOPS/W. We have found experimental evidence for this in the study of massively parallel array processors, mainly dedicated to image processing. Their distributed architecture reduces the energy overhead dedicated to data handling, thus resulting in a power efficient implementation.**

***Index Terms*—Parallel processing, Cellular Processing Array, Multicore processing, Computational efficiency**

I. Introduction

Distributing tasks between several elementary processors working in parallel speeds up processing. This is constrained, however, by the degree of parallelization that can be achieved [1]. Amdahl argued that this limitation favors the use of a single-processor system in order to achieve large computing efforts [2]. However, in cases in which the amount work grows beyond tractability, parallel is the only alternative to operate on a large amount of data in a certain amount of time [3]. In addition to this, there is a less intuitive result that is related with energy efficiency [4], which is that hardware parallelization renders more efficient implementations. While power consumption scales linearly with the operating frequency, computing power measured in MIPS (million instructions per second) does not. This happens because computing power does not depend exclusively on the processor performance, but on other factors that do not scale with the clock, like memory operations or data transmission. Distributing resources is the key to overcome this limitation. By placing the necessary resources close to the processing elements, e. g. sensors, memory cells, the overhead introduced by data handling is reduced. This architectural intervention represents an adaptation to the nature of the signal. Flexibility is traded for efficiency.

While Amdahl's work concentrated in performance, there are scenarios in which power efficiency is the main driver, for instance image processing in embedded vision systems. In this study we will consider single- and multicore processor chips, GPUs and different types of SIMD (single-instruction multiple-data) arrays containing from tens to hundreds of processing elements (PE). In all cases, they will be single-chip devices.

II. A Survey on Power-Efficient Processors

In order to validate the hypothesis, we have collected data from processors designed, fabricated and tested from 2003 to 2013. A total of 65 processors have been considered for this analysis. The processors' data, together with their bibliographic references and a complete table in Excel format can be downloaded from `http://www2.imse-cnm.csic.es/mondego/public/`. Most of them employ more than one single core and are oriented to portable applications, being therefore designed for power efficiency. We have also included some analog and mixed-signal array processors, although their heterogeneous design style, operating principles and throughput estimation, can introduce a noticeable distortion of the results.

In order to make these processors comparable, we will adopt the terminology employed in [1]. Therefore, each particular multicore processor chip contains n base core equivalents (BCEs), that will be the elementary processor. These resources could be employed to implement n of the simplest cores —one BCE each—, working in parallel, or to implement a single processor with all the n BCE resources. Between these extremes, multicore chips with N_{proc} cores using r BCE resources each. It is easily seen that $N_{\text{proc}} = n/r$.

Now, the silicon area, A, occupied by a circuit is related with its complexity. This figure needs to be normalized by λ^2, as the chips in the survey have been fabricated with different resolutions. In order to establish a common reference, the normalized area of the BCE, A_0, needs to be defined. We have divided all the normalized areas of the chips in the table by N_{proc}, and have chosen the smallest value. Notice that this is a normalizing factor, so it will not distort the results. Each chip will be characterized by a pair of values: the total number of resources n and the resources of each individual core:

$$(n, r) = \left(\frac{A/\lambda^2}{A_0}, \frac{n}{N_{\text{proc}}} \right) \tag{1}$$

III. Performance vs. complexity

According to Pollack's rule, performance scales with the square root of any increase in complexity [5]. Based on this, the throughput of a processor composed of n/r cores with r BCEs each is given by:

$$G(n, r) = \frac{n}{\sqrt{r}} G_0 \tag{2}$$

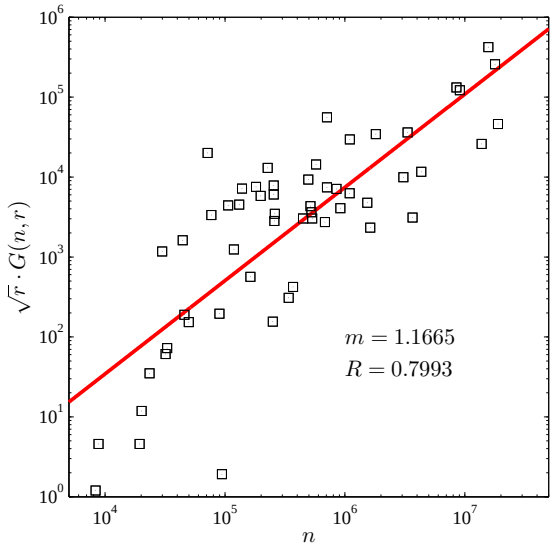

Fig. 1. Pollack's rule: GOPS vs. n

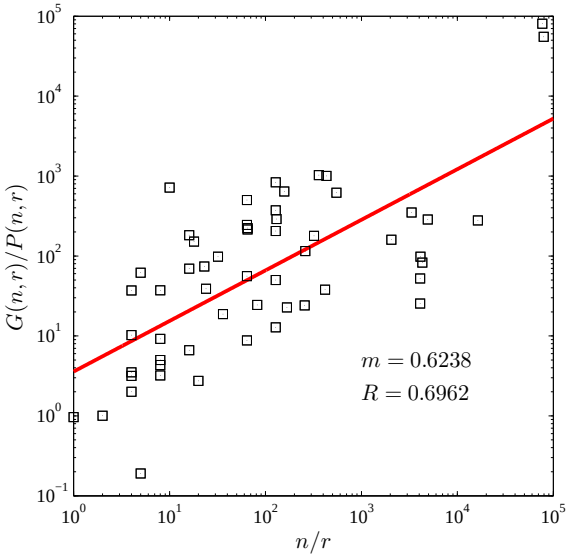

Fig. 2. Power efficiency vs. number of parallel processors

where G_0 is the performance of the BCE. Notice that Eq. (2) encompasses the cases of a single processor using the n BCE resources ($r = n$) and of n elementary one-BCE-processors working in parallel ($r = 1$). If we represent $\sqrt{r}G(n,r)$ vs. n we should obtain a straight line. Fig. 1 represents this magnitude for the 56 digital processors in the survey. The least-squares regression line has a slope of $m = 1.1665$, being $R = 0.7993$ the correlation coefficient. This confirms the linear dependence of $\sqrt{r}G(n,r)$ with n. Given the relative disparity of design style, system-on-chip architecture and application field, the alignment found can be enough to identify a trend.

IV. Power efficiency vs. No. of processors

If we plot power efficiency, $G(n,r)/P(n,r)$, vs. the number of processors working in parallel (n/r) in each processor (Fig. 2), it can be observed that it grows as roughly:

$$\frac{G(n,r)}{P(n,r)} \sim \left(\frac{n}{r}\right)^{\frac{2}{3}} \tag{3}$$

Given the different design principles, architectures and internal organization of resources employed at the different processors, the correlation found ($R = 0.6962$) denotes a trend that is: the more processors working in parallel the higher power efficiency. Notice that increasing the number of resources in order to build one single but more complex sequential processor ($r = n$) does not have an incidence in power efficiency.

With respect to the 9 analog and mixed-signal array processors in the list, it is not possible to find an elementary processor to establish a comparison. Area is the closest estimation of circuit complexity that can be employed. However, the trend on performance vs. area is highly correlated ($R = 0.8170$), although the relation responds to a different model. No power efficiency trend has been identified, but area efficiency vs. N_{proc} is highly correlated as well ($R = 0.9269$).

V. Conclusions

It is an observable trend that increasing the degree of parallelism in hardware has a positive incidence not only in performance but also in power efficiency. There must be a fundamental reason for that, given that processors developed by following different design strategies and internal organization of resources, follow the common behavior displayed by experimental data. In analog and mixed-signal processors, performance increases with N_{proc} but the disparity in the design approaches end up in highly uncorrelated data concerning power efficiency. Area efficiency however grows —it needs to— when more processors are packed into the same chip.

Acknowledgments

This work has been funded by the Spanish MINECO through project TEC2015-66878-C3-1-R and CDTI through IPC- 20111009 (both co-funded by the European Regional Development Fund ERDF/FEDER), Junta de Andalucía's CEICE through TIC 2338-2013 CEICE, and the Office of Naval Research (USA) through grant No. N000141410355.

References

[1] M. Hill and M. Marty, "Amdahl's law in the multicore era," *Computer*, vol. 41, no. 7, pp. 33 –38, July 2008.

[2] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proc. of the Spring Joint Computer Conference (AFIPS)*, 1967, pp. 483–485.

[3] J. L. Gustafson, "Reevaluating Amdahl's law," *Communications of the ACM*, vol. 31, no. 5, pp. 532–533, May 1988.

[4] S. Moreno-Londono and J. Pineda de Gyvez, "Extending Amdahl's law for energy-efficiency," in *Int. Conf. on Energy Aware Computing (ICEAC)*, Dec. 2010, pp. 1–4.

[5] S. Borkar, "Thousand core chips: a technology perspective," in *Proceedings of the 44th annual Design Automation Conference (DAC)*, 2007, pp. 746–749.