# A Performance-Energy Model to Evaluate Single Thread Execution Acceleration

Seung Hun Kim, Dohoon Kim, Changmin Lee, Won Seob Jeong, *Student Member, IEEE*, Won Woo Ro, *Member, IEEE*, and Jean-Luc Gaudiot, *Fellow, IEEE*

**Abstract**—It is well known that the cost of executing the sequential portion of a program will limit and sometimes even eclipse the gains brought by processing in parallel the rest of the program. This means that serious consideration should be brought to bear on accelerating the execution of this unavoidable sequential part. Such acceleration can be done by boosting the operating frequency in a symmetric multicore processor. In this paper, we derive a performance and power model to describe the implications of this approach. From our model, we show that the ratio of performance over energy during the sequential part improves with an increase in the number of cores. In addition, we demonstrate how to determine with the proposed model the optimal frequency boosting ratio which maximizes energy efficiency.

**Index Terms**—Performance modeling, Multiprocessor systems, Energy-aware systems

## 1 Introduction

While processing in parallel the multiple threads of an otherwise single process, the well-known Amdahl's law has shown that the speedup which could be delivered would be restricted by the cost of executing of the "unavoidable" sequential portion in the program [1] (*i.e.,* that which cannot be logically parallelized). Therefore, improving the performance of this sequential part is crucial if we are to ever improve the overall performance of parallel processors (such as modern multicore architectures).

One of the possible approaches to accelerate the sequential part is to increase the operating frequency of the specific core for the sequential part. For example, Intel *Turbo Boost* technology enables transient overclocking for a dedicated core if the other cores are in an idle state [2]. This may be a reasonable technique since only one core is needed for the sequential part. However, the method inevitably causes increased energy consumption due to the increased operating frequency. Therefore, the trade-off between performance and energy should be carefully considered. We therefore present here a performance and power model and show the energy efficiency through the proposed modeling method. Consequently, we demonstrate that the optimal boosting ratio which maximizes the efficiency can be determined. In addition, we develop our theoretical model to show the energy efficiency with different design styles of a symmetric multicore processor.

For the purposes of modeling and measurement we refer to previous studies [3], [4]. However, unlike prior research, our work focuses on the energy efficiency of the sequential part acceleration. We evaluate the effectiveness of the frequency boosting technique in terms of performance and power by comparing it to a baseline system with no acceleration of the sequential portions. Our analysis of the results allows us to propose guidelines to design symmetric multicore processors with an energy-efficient acceleration of the sequential portions of the application programs.

- *S.H. Kim, D. Kim, C. Lee, W.S. Jeong, and W.W. Ro are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea. E-mail: kseunghun@gmail.com, {dohoon.kim, exahz, ws.jeong, wro}@yonsei.ac.kr*
- *J.-L. Gaudiot is with the Department of Electrical Engineering and Computer Science, University of California, Irvine, California, U.S.A. E-mail: gaudiot@uci.edu*

## 2 Performance and Power Models for the Acceleration

### 2.1 Related Work

Hill and Marty presented a performance model of multicore processors using Amdahl's law [3]. They proposed the concept of Base Core Equivalent (BCE) resource to predict the performance according to processor design styles. Their proposed models help to provide a proper approach to design multicore processors and motivated follow-up work. For example, Woo and Lee studied the power model of multicore architecture for energy efficient parallel processing [4]. In the study, they defined two core types as *energy-efficient small core* and *performance-enhanced large core* and modeled performance and power of each core type to follow Hill and Marty's work [3]. Also, Sun and Chen introduced a more optimistic performance model in multicore scalability than that of Hill and Marty [5]. In Sun and Chen's study, the fixed-time speedup model which is proposed by Gustafson [6] was used to show the scalability of symmetric multicore architectures.

In addition to the performance and power analysis, other studies presented mathematical approaches to guide processor design and operation policy. For multicore processor utilization, Cho and Melhem presented the relationship between the speedup and energy consumption in parallel processors. Also, they provided a method which minimizes energy consumption without any significant performance degradation [7]. Zidenberg *et al.* proposed the MultiAmdahl framework to show an optimal resource allocation method such as power and area for heterogenous processor [8] and Morad *et al.* extended the work for Multi-Accelerator architectures [9].

### 2.2 Effectiveness of the Acceleration

In this section, we present the performance and power model of the sequential part acceleration in a parallel environment. We focus on the acceleration method based on operating frequency boosting, and thus a new variable $B$ is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCA.2014.2368144, IEEE Computer Architecture Letters



introduced to represent the ratio of the increased frequency to the baseline. Generally, the performance of each core in the processor is defined as a unit performance in Amdahl's law. Therefore, if the acceleration enables $bst\_perf(B)$ times improvement in the performance of the sequential portion, the achievable maximum speedup ($Perf$) can be obtained as follows:

$$Perf = \frac{1}{(\frac{s}{bst\_perf(B)}) + \frac{1-s}{n}} \quad (1)$$

where $s$ is the serial portion and $n$ is the number of available cores.

The function $bst\_perf(B)$ is the performance improvement according to the operating frequency where we assume that each core converts 65% of increased frequency into an improvement in performance (a reasonable assumption also made by Sprangle and Carmean [10]). In fact, any performance improvement is not directly proportional to the operating frequency since various other factors such as data access latency and cache misses come into play. The assumption that we have borrowed from their work means that 65% of the total execution time is affected by the operating frequency and the other 35% is not related to the frequency. Consequently, the function can be expressed as

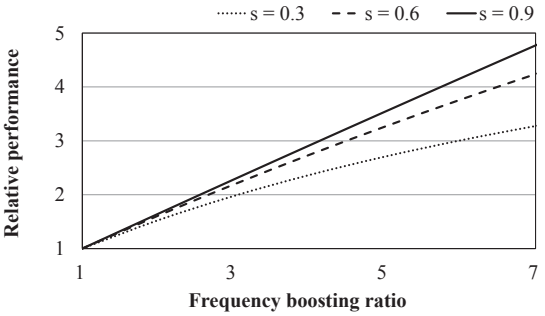
$$bst\_perf(B) = 0.65 \cdot B + 0.35 \quad (2)$$

Also, we can estimate the average power consumption of the acceleration by using Woo and Lee's formula [4]; we assume that one core consumes a power of 1 in the active state and $k$ in the idle state ($0 \leq k \leq 1$)[1]. When one core is accelerated with a frequency boosted by a multiplicative factor $B$, we follow standard DVS literature [7] and assume that power consumption increases by a factor $B^3$. Consequently, the average power consumption ($W$) of the sequential part acceleration becomes

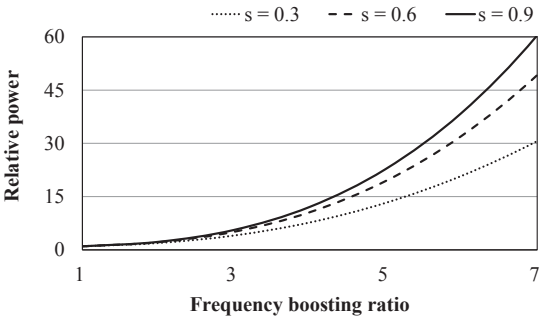
$$W = \left\{ \frac{s}{bst\_perf(B)} \cdot (B^3 + (n-1) \cdot k) + (1-s) \right\} \cdot (Perf) \quad (3)$$

Equation (1) (with the inclusion of (2)) and (3) give us the performance and power respectively when some of the execution may take place in parallel. This model takes into account the acceleration of the sequential part. Fig. 1 uses these equations to show how performance and power vary with the frequency boosting ratio on a 16 core machine, with the serial portions set to 30%, 60%, and 90%, respectively. The modeling results are normalized to a baseline model where processing in parallel takes place, but without acceleration of the sequential portion. Fig. 1a shows that boosting the frequency will yield a meaningful speedup and it also shows that the benefits from the acceleration scheme increase with a larger proportion of the serial portion. However, the relative increase in power consumption is much higher than that in performance; there are indications that power consumption will continue growing exponentially with increases in the boosting ratio while the performance improvement appears linear at best and may even saturate at some point. This trend strongly implies that there should be an upper limit for the frequency boosting ratio ($B$) to achieve energy efficient acceleration; we provide a detailed analysis in the next section.

1. In the rest of the paper, $k$ is assumed to be 0.3, although, incidentally, we have found no remarkable changes to our results with different values of $k$.




(a) Relative performance of the acceleration




(b) Relative average power of the acceleration

Fig. 1. Performance and power of the sequential part acceleration

## 3 Energy Efficient Acceleration

### 3.1 Energy Efficiency

In order to quantitatively evaluate energy efficiency, we use the *performance per joule* metric presented by Woo and Lee [4]. The metric represents the achievable performance improvement for a given amount of energy.

$$\frac{Perfomance}{Joule} = \frac{\frac{1}{T_b}}{T_b \cdot W} = \frac{1}{{T_b}^2 \cdot W} = \frac{Perf^2}{W} \quad (4)$$

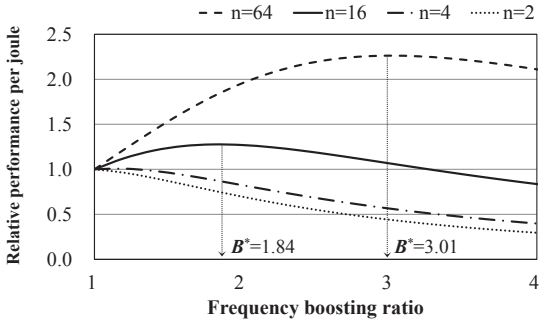
where $T_b$ is the reduced execution time using the acceleration; therefore, $\frac{1}{T_b}$ is ($Perf$).

From (4), the efficiency of the sequential part acceleration with regard to the frequency boosting ratio can be shown in Fig. 2. The number of cores are set to 2, 4, 16, and 64 for two cases of serial portion: 30% and 90%. Also, the results are normalized to the case which has no acceleration; normalization is respectively done for each number of cores. That is, we focus our evaluation on determining the energy efficiency of the acceleration according to the value of the boosting ratio for a given multicore architecture.

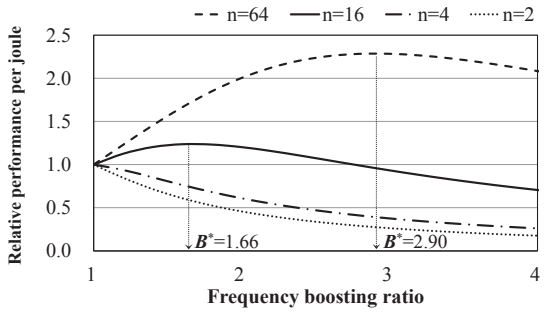
Fig. 2a and Fig. 2b show the performance results with small and large serial portions, respectively. As is evident from the figures, the acceleration provides improved energy efficiency for a certain range of $B$ values if the number of cores is relatively large. On the other hand, if the number of cores is small, increasing the operating frequency for the acceleration degrades the energy efficiency. In fact, the amount of improvement increases as the number of cores increases with both the small and large serial portions. The main reason is the high power consumption while execution in parallel is taking place. As mentioned by Woo and Lee, the power required to complete a job of a given

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCA.2014.2368144, IEEE Computer Architecture Letters






(a) Results for a small serial portion (s=0.3)




(b) Results for a large serial portion (s=0.9)

Fig. 2. Performance per joule of the sequential part acceleration

size grows with the number of cores [4]. For our evaluation, the influence of the increased power consumption due to the frequency boosting diminishes with an increase in the number of cores. Still, the acceleration causes a decrease in the overall execution time. Therefore, the acceleration of the sequential part will yield better energy efficiency if the number of cores and the frequency boosting ratio are carefully selected.

### 3.2 Optimal Boosting Ratio

As shown in Fig. 2, there is a maximum value for the *performance per joule* metric for the varying frequency boosting ratio at a given number of cores and for a specific serial portion. In other words, a boosting ratio optimal with respect to energy efficiency can be determined by finding the value $B^*$ which maximizes the result of (4). Therefore, $B^*$ can be obtained by deriving (4) with respect to $B$. From the derivative, we found the optimal boosting ratio according to the serial portion and the number of cores. The results are shown in Fig. 3.

The results provide useful insight in the design of symmetric multicore processors regarding the acceleration of sequential portions in programs. First of all, the ability to estimate the serial portion of a program will be beneficial to decide on the optimal boosting ratio for energy efficiency. As shown in Fig. 3, the $B^*$ values are highly dependent on the serial portion when that portion is small. Hence, the boosting ratio can be determined for each program using an approximation of the size of the serial portion and the number of cores; profiling for the threads of applications at run-time would be an example for the approximation method.

Second, when a precise estimation of the serial portion is difficult to obtain, we can use the minimum $B^*$ as a reference boosting ratio which is close to every $B^*$ value for

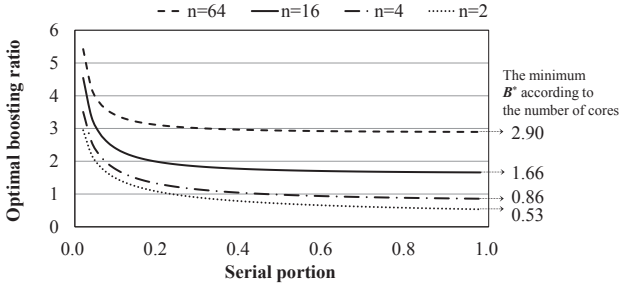



Fig. 3. Optimal frequency boosting ratio maximizing energy efficiency

a wide range of serial portions. As shown in Fig. 3, the slope of the $B^*$ values decreases as the serial portion increases. Therefore, having the reference value as the minimum $B^*$ sounds reasonable when the serial portion is more than 20% of the program. The minimum $B^*$ values according to the number of cores are shown on the right side of the graph.

In fact, the usefulness of the reference boosting ratio can be demonstrated with the results shown in Fig. 2. For example in the 16 core case, the minimum $B^*$ is 1.66 which still works well both for a small serial portion (Fig. 2a) and a large serial portion in (Fig. 2b). However, if the reference value is determined as a large (*e.g.*, $B = 4.0$), it can cause performance to degrade in both cases.

### 3.3 Chip Cost Equivalent Model for the Acceleration

We have shown the performance and power model of the sequential part acceleration and derived the optimal frequency boosting ratio for a given number of cores. However, we have not considered the variations in a symmetric multicore design style over a fixed chip area. In fact, the design style is an important factor of the performance and power analysis [4]. For a restricted chip area, a processor can be designed to have fewer cores (but be more powerful) or to have more cores (but be less powerful). While the former design style has advantages as far as the sequential execution is concerned, the later has advantages while parallel execution is taking place. Also, the power consumption of a single core is larger with the former. Considering these facts, we can evaluate the energy efficiency of the acceleration according to single core performance for the fixed area of a symmetric multicore processor.

We introduce two new variables $s_p$ and $s_w$ to represent the scaled performance and power with regard to the baseline where both variables are 1. We assume that the power consumption is proportional to the chip area if the other conditions such as operating voltage and frequency remain the same. The values of the variables are decided according to chip area. More specifically, if the number of transistors in a core increases $C$ times with regard to the baseline, $s_p$ and $s_w$ become $\sqrt{C}$ and $C$, respectively [11]. Also, the number of cores in the processor is proportional to $1/C$ in this case. Finally, the performance and average power model that are shown in (1) and (3) can be rewritten as

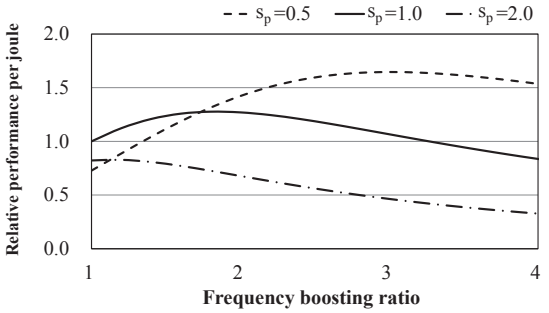
$$Perf_{CE} = s_p \cdot Perf \tag{5}$$

$$W_{CE} = s_w \cdot W \tag{6}$$

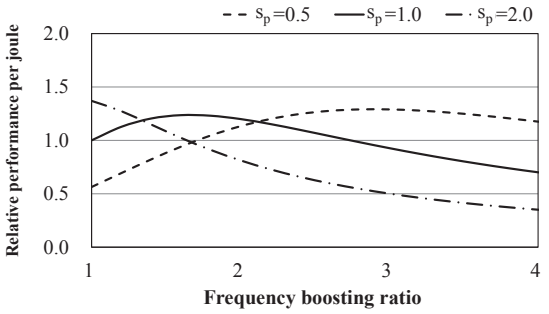
Now energy efficiency of the acceleration can be compared for different symmetric multicore design styles. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCA.2014.2368144, IEEE Computer Architecture Letters






(a) Results for small serial portion (s=0.3)




(b) Results for large serial portion (s=0.9)

Fig. 4. Performance per joule of the sequential part acceleration with the equivalent chip cost

baseline case is defined as 16 cores and the results are shown in Fig. 4. The design styles are assumed as the lower and higher performance of a single core compared to the baseline, and thus the values of $s_p$ are set as 0.5 and 2.0, respectively; therefore, the numbers of cores are 64 and 4 in each case. The results are normalized to the baseline that has normal operating frequency.

Fig. 4a shows the performance results when the serial portion is small. As shown in the figure, the energy efficiency of $s_p = 0.5$ case shows the best performance when the frequency boosting ratio is larger than 1.73. Although the performance of a single core is diminished by $0.5$, the time spent in the parallel part becomes smaller than the baseline since the total number of cores is increased. Therefore, the acceleration yields improved energy efficiency by allowing faster execution of the sequential part. On the contrary, the observed efficiency of $s_p = 2.0$ is always lower than that of the baseline and the efficiency worsens as the boosting ratio increases. These results imply that the dominant factor if we are to provide an energy efficient acceleration is the number of cores rather than the performance of each core when the serial portion is small.

When the serial portion is large (Fig. 4b), the energy efficiency is lowest in the case of a larger number of cores (*i.e.*, $s_p = 0.5$) if there is no acceleration. Lee and Woo also observed similar results under similar conditions including serial portion, $s_p$, and $s_w$ [4]. For this reason, they concluded that increasing the number of cores has disadvantages with large serials portion in terms of energy efficiency. However, their conclusion is valid only if no consideration is made for the sequential part acceleration. As shown in the figure, the efficiency increasing range for the frequency boosting ratio is widest when $s_p = 0.5$. For this reason, the results of the baseline and of the case $s_p = 0.5$ are reversed after a certain value of the boosting ratio; based on our model, the value is near to $1.68$. In addition, the attainable maximum efficiency of the $s_p = 0.5$ is larger than that of the baseline. That is, increasing the number of cores is the most important factor to design an energy efficient sequential part acceleration in a symmetric multicore architecture.

## 4 Conclusions

Accelerating the sequential part of a program is a promising approach to improve overall performance in parallel processors. For this acceleration, we investigated the operating frequency boosting method in a symmetric multicore processor and showed the expected performance improvement and the power consumption of the method based on Amdahl's law. Also, we have applied the proposed energy efficiency model to different symmetric multicore design styles. From the results, we found that the energy efficiency of the acceleration increases with the number of cores and an optimal frequency boosting ratio can be determined. We will further develop the hardware architecture for the acceleration considering the proposed modeling results and the distinguishing characteristics of the various parallel programs.

## Acknowledgments

This material is based upon work supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2010-0013202) and by the National Science Foundation, under award CCF-1439165. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

## References

[1] G. M. Amdahl, "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities," in *Proceedings of the American Federation of Information Processing Societies (AFIPS)*, Apr. 1967, pp. 483–485.

[2] Intel Corporation, "Intel Server Board and System Products Update on Intel Turbo Boost Technology Support with Low Power Intel Xeon Processor 3400/5500/5600 Series," *Whitepaper*, Apr. 2010.

[3] M. Hill and M. Marty, "Amdahl's Law in the Multicore Era," *Computer*, vol. 41, no. 7, pp. 33–38, July 2008.

[4] D. H. Woo and H.-H. Lee, "Extending Amdahl's Law for Energy-Efficient Computing in the Many-Core Era," *Computer*, vol. 41, no. 12, pp. 24–31, Dec. 2008.

[5] X.-H. Sun and Y. Chen, "Reevaluating Amdahl's Law in the Multicore Era," *Journal of Parallel and Distributed Computing*, vol. 70, no. 2, pp. 183–188, Feb. 2010.

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

[7] S. Cho and R. Melhem, "Corollaries to Amdahl's Law for Energy," *Computer Architecture Letters, IEEE*, vol. 7, no. 1, pp. 25–28, Jan. 2008.

[8] T. Zidenberg, I. Keslassy, and U. Weiser, "MultiAmdahl: How Should I Divide My Heterogenous Chip?" *Computer Architecture Letters, IEEE*, vol. 11, no. 2, pp. 65–68, Apr. 2012.

[9] A. Morad, T. Morad, L. Yavits, R. Ginosar, and U. Weiser, "Generalized MultiAmdahl: Optimization of Heterogeneous Multi-Accelerator SoC," *Computer Architecture Letters, IEEE*, no. 99, pp. 1–1, 2012. http://dx.doi.org/10.1109/L-CA.2012.34.

[10] E. Sprangle and D. Carmean, "Increasing Processor Performance by Implementing Deeper Pipelines," in *Proceedings of the 29th Annual International Symposium on Computer Architecture (ISCA)*, May 2002, pp. 25–34.

[11] S. Borkar, "Thousand Core Chips: A Technology Perspective," in *Proceedings of the 44th Annual Design Automation Conference (DAC)*, June 2007, pp. 746–749.