Information Processing Letters 111 (2011) 1092–1095

Contents lists available at SciVerse ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

What Hill–Marty model learn from and break through Amdahl's law?

Erlin Yao *, Yungang Bao, Mingyu Chen

State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 30 November 2010
Received in revised form 13 September 2011
Accepted 14 September 2011
Available online 19 September 2011
Communicated by A.A. Bertossi

Keywords:
Parallel processing
Multicore
Scalability
Amdahl's law
Hill–Marty model

ABSTRACT

Chip multiprocessors (CMPs) or multicores are emerging as the dominant computing platform. Recently, Hill and Marty developed the model which augmenting Amdahl's law to the multicore hardware, and this model has received great attention. However, there are still the following fundamental problems in the perspective of theory remains unsolved, such as: How general is the observation in Hill–Marty model, does it only hold for some specific architectures? As a corollary of Amdahl's law, what does Hill–Marty model learn from and break through Amdahl's law? This paper investigates an analytical and quantitative analysis to these problems, the obtained results could provide computer architects with a better understanding of multicore scalability.

1. Introduction

The multicore scalability is an important problem for the future generations of chips with multiple processor cores. Recently, Hill and Marty [6] augmented Amdahl's law to multicore hardware by constructing a cost model for the number and performance of cores that the chip can support. They observed that obtaining optimal multicore performance will require further research in both extracting more parallelism and making sequential cores faster. They indicated that in addition to working on the parallelization of applications, researchers should also pay attention to investigating methods of speeding sequential performance even if they appear locally inefficient, since these methods can be globally efficient as they reduce the sequential phase, which is the bottleneck in Amdahl's framework. This work (Hill–Marty model) has received much attention in the parallel computing and multicore area [2,11]. The related applications of this model have varied from the instruction level [10], operating systems [3], to the TRIPS computer system [4] and multicore algorithm designs [9].

Of course, Hill and Marty's model has deficiency. Indicated by Hill and Marty themselves, this model ignored the important effects of dynamic and static power, as well as on- and off-chip memory system and interconnect design, etc. [6]. And the communication cost between cores is also an important factor not considered in their model. However, in this paper, we will not discuss how to incorporate these missing factors to this model or how to apply this model to real applications.

On the other hand, we will study the following fundamental problems from theoretical perspective, such as: Is the observation in Hill–Marty model specific or general for multicore architectures? And since Hill–Marty model is only a corollary of Amdahl's law, what is the breakthrough of this model to Amdahl's law? And what does this model learn from Amdahl's framework?

2. Amdahl's law and Hill–Marty model

More than four decades ago, Gene Amdahl defined his law for the special case of using p processors in parallel when he argued for the single-processor approach's validity for achieving large-scale computing capabilities [1]. He used a limit argument to assume that a fraction f of a pro-

* Corresponding author.
E-mail addresses: yaoerlin@ict.ac.cn (E. Yao), baoyg@ict.ac.cn (Y. Bao), cmy@ict.ac.cn (M. Chen).

0020-0190/$ – see front matter
doi:10.1016/j.ipl.2011.09.009

gram's execution time was infinitely parallelizable with no scheduling overhead, while the remaining fraction, $1-f$, was totally sequential. Without presenting an equation, he noted that the speedup on p processors is governed by:

$$Speedup_{\text{parallel}}(f,p)=\frac{1}{1-f+f/p}.$$

Despite its simplicity, Amdahl's law applies broadly and gives important insights such as: (i) Attack the common case: When f is small, optimization will have little effect; (ii) The aspects you ignore also limit speedup: Even if p approaches infinity, the speedup is bounded by $1/(1-f)$.

Hill and Marty augmented Amdahl's law to multicore hardware by constructing a cost model for the number and performance of cores that the chip can support [6]. They first assumed that a multicore chip of given size and technology generation can contain at most n base core equivalents (BCE) (where a single BCE implements the baseline core). Second, they assumed that architects can use the resources of multiple BCEs to create a core with greater sequential performance. If the performance of a single-BCE core is 1, they assumed that architects can expend the resources of r BCEs to create a powerful core with sequential performance $perf(r)$, where $perf$ is an increasing function and satisfies $1<perf(r)<r$. According to the cost model, they classify the architecture of multicore chips into three categories: symmetric, asymmetric and dynamic multicore chips.

In this paper we choose the recently popular asymmetric architecture as an example. In an asymmetric multicore chip, there is one big core (consuming more BCEs) which is more powerful than the others (all consuming one-BCE resource). With a resource budget of n BCEs, an asymmetric multicore chip can have $1+n-r$ cores because the single larger core consumes r $(1\leqslant r\leqslant n)$ resources and leaves $n-r$ resources for the one-BCE cores. This chip uses the one core with more resources to execute sequentially at performance $perf(r)$. In the parallel fraction, however, it gets performance $perf(r)$ from the large core and performance 1 from each of the $n-r$ base cores. Overall, according to Amdahl's law, the speedup of an asymmetric multicore chip (relative to using one single-BCE core) is [6]:

$$Speedup_{\text{asymmetric}}(f,n,r)=\frac{1}{\frac{1-f}{perf(r)}+\frac{f}{perf(r)+n-r}}.$$

In this paper we assume that the parallel fraction f is not in the extremes ($f=0$ or $f=1$), which can be easily treated separately. So it holds that $0<f<1$.

3. How general is the observation in Hill–Marty model?

Note that in Hill–Marty's speedup formula, the parallel fraction f and the performance function $perf$ are supposed to be known beforehand. The number of resources n is also fixed when the multicore chip's size and technology generation are given. So the only leaving variable parameter is r. From the definition of r, it can be seen that $1\leqslant r\leqslant n$. One of the main purpose of their work is to determine the optimal r_0 in the perspective of speedup. According to the definition of r_0 in asymmetric architecture, it is clear that $r_0=1$ means that the optimal architecture is all the cores are base cores, which indicates that we should exploit the parallelism as more as possible and do not need to make sequential cores faster; however, $r_0=n$ means that we should build a chip with only one big core including all the resources, and this indicates that the parallelism is not important but all the attempt should be paid to make sequential cores faster.

However, Hill and Marty observed that generally the optimal speedup will not occur at the extremes ($r_0=1$ or $r_0=n$), but between the extremes ($1<r_0<n$). So they arrived at the observation that "obtaining optimal multicore performance will require further research in both extracting more parallelism and making sequential cores faster" [6]. Note that in all their graphs in [6], Hill and Marty assumed that $perf(r)=r^{0.5}$, which is a very specific function. Nonetheless, Hill and Marty also tried other similar functions (for example, $r^{2/3}$), but found no important changes to their results. They also stated that their equations allow $perf$ to be any arbitrary function [6]. In their website for this project [13], they also provided a program in which the performance function can be chose as the combination of any elementary mathematical functions, and the number of resources n can be selected as any reasonable number.

According to the speedup formula of asymmetric multicore, if the parallel fraction f is fixed, then it is clear that the choice of performance function $perf$ and the number n is critical to the property of the optimal r_0. According to Moore's law, the number of BCE resources n will increase continually as technology develops. Note that, Hill and Marty's observation cannot be generalized to any arbitrary performance function and reasonable n only through experiments on some specific functions and numbers n using computer program. To validate whether their observation holds in a general framework, theoretical proof is needed.

In [12], we had theoretically proved that Hill and Marty's observation holds for any performance function in the form of $perf(r)=r^c$ with $0<c<1$. However, in real applications, the performance function might be more complex than r^c. For example, it can be: $perf(r)=0.7r^{0.5}+0.2r^{1/3}+0.1$. Although theoretically the performance function $perf$ can be any reasonable function, it should also satisfy the following practical restrictions: (i) $perf(1)=1$; (ii) $perf(r)<r$; (iii) $perf(r)$ is an increasing function of r; (iv) $\lim_{r\to\infty} perf'(r)=0$. The first three conditions are clear. The fourth condition states that $perf$ should be like a saturated function, which means that when the number of resources including in a core is big enough, the performance gain of increasing resources continually will be very small. Then we can prove the following result.

Claim 1. *If the number of resource budget including in a multicore chip increases continually as technology develops, then Hill and Marty's observation holds for any reasonable performance function satisfying practical restrictions.*

Proof. We view the speedup as a function of r:

$$S(r)=Speedup_{\text{asymmetric}}(f,n,r)=\frac{1}{\frac{1-f}{p(r)}+\frac{f}{p(r)+n-r}}.$$

Since n will increase continually, we only need to prove that $\lim_{n\to\infty} S(1)$ and $\lim_{n\to\infty} S(n)$ will not be the maximal speedup. According to the rule of first derivative,

$$S'(r) = S^2(r)\left[\frac{(1-f)p'(r)}{p^2(r)} + \frac{f(p'(r)-1)}{(p(r)+n-r)^2}\right].$$

Since $p(1) = 1$, it holds that

$$\lim_{n\to\infty} S'(1) = \lim_{n\to\infty} S^2(1)\left[(1-f)p'(1) + \frac{f(p'(1)-1)}{n^2}\right] = \frac{p'(1)}{1-f} > 0.$$

Then it is clear that the maximum of speedup will not be achieved at $r = 1$. And it holds that:

$$\lim_{n\to\infty} S'(n) = \lim_{n\to\infty} \left(p'(n) - f\right) = -f < 0.$$

Then it can be seen that the maximal speedup will not be achieved at $r = n$ either. □

Claim 1 indicates that as technology develops according to Moore's law, for any practical performance function, obtaining optimal multicore performance will require further research in both extracting more parallelism and making sequential cores faster.

4. What is the breakthrough of Hill–Marty model to Amdahl's law?

Recall that in Amdahl's law, even if the number of processors p approaches infinity, the speedup is bounded by $1/(1-f)$. In Hill and Marty's model, if the number of resource budget n approaches infinity, will the speedup also be limited?

Claim 2. *If the number of resource budget n increases continually as technology develops, then the optimal speedup in Hill–Marty model will not be limited by a constant.*

Proof. Note that for any reasonable performance function, $\lim_{n\to\infty} perf(n) = \infty$ should be satisfied. If we consider the speedup at $r = n$, then it is clear that

$$\lim_{n\to\infty} Speedup(r = n) = \lim_{n\to\infty} perf(n) = \infty.$$

So it can be seen that in Hill–Marty model, the optimal speedup is not limited by a constant when the number of resources increases continually as technology develops according to Moore's law. □

In Claim 2 it seems puzzling that the performance could be unbounded given unbounded resources. Here we give a practical example: If the constraining resource on performance is power, then given unlimited power (and heat dissipation capabilities), it should be possible to achieve unlimited performance. Since given unlimited power and cooling, you can just increase the processor's clock frequency to infinity leading to infinite performance gains.

Since Hill–Marty model is only a corollary of Amdahl's model, this introduced unlimited speedup can be viewed as the biggest breakthrough of Hill–Marty model to Amdahl's framework. It can be seen that the breakthrough of speedup is actually caused by the introducing of the performance function *perf*. Although, currently hardware designers can't build cores that achieve arbitrary high performance by adding more resources. However, Hill and Marty's work provides an idea to overcome the limitation of Amdahl's framework – making sequential cores faster. Researchers should investigate methods of speeding sequential performance even if they appear locally inefficient (for example, $perf(r) = r^{0.5}$). These methods can be globally efficient as they reduce the sequential phase, which is the bottleneck in Amdahl's framework.

5. What does Hill–Marty model learn from Amdahl's law?

Although Hill and Marty's work overcomes the most significant limitation of Amdahl's framework, since it is a corollary of Amdahl's model, it must have inherited some properties of Amdahl's model. Among these properties, which one is the most important and inherent under Amdahl's framework? The following result can be given.

Claim 3. *In Hill–Marty model, the architecture of including all the BCE resources in one single core is always sub-optimal, which reproduces Amdahl's argument on the validity of the single-processor approach to achieving large-scale computing capabilities.*

Proof. Suppose the optimal architecture is $r = r_0$, then for any $1 \leqslant r \leqslant n$, it holds that

$$Speedup(f, n, r) \leqslant Speedup(f, n, r_0).$$

And it is clear that

$$Speedup(f, n, r = n) = perf(n),$$

so we have

$$perf(n) \leqslant Speedup(f, n, r_0).$$

On the other hand, it holds that

$$Speedup(f, n, r) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f}{perf(r)+n-r}} < \frac{1}{\frac{1-f}{perf(r)}} = \frac{perf(r)}{1-f}.$$

Since *perf* is an increasing function, so $perf(r) \leqslant perf(n)$. Then for any $1 \leqslant r \leqslant n$, we have

$$Speedup(f, n, r) < perf(n)/(1-f).$$

Then it is clear that

$$Speedup(f, n, r_0) < perf(n)/(1-f).$$

So it holds that

$$perf(n) \leqslant Speedup(f, n, r_0) < perf(n)/(1-f).$$

This indicates that the optimal performance is always within $1/(1-f)$ times of the performance of the architecture including all the BCE resources in one single core. □

Claim 3 states that although in Hill and Marty's model, obtaining optimal multicore performance will require both extracting more parallelism and making sequential cores faster, however, in common cases it only requires limited parallel computing. This reproduces Amdahl's argument, and can be viewed as the most important and inherent property of Hill–Marty model originated from Amdahl's framework.

Note that, Amdahl's law is also known as the fixed-size speedup model, which assumed that the problem size, and so the parallelization fraction f, has not increased as the computing power scales up. However, since more computing resources might advantageously allow greater parallelism from larger problem size, future works should also investigate extending the Gustafson's law (fixed-time speedup model) [5] and Sun–Ni's law (memory-bounded speedup model) [7,8] in the multicore era.

6. Conclusion

This paper could provide computer architects with a better and quantitative understanding of Hill and Marty's work on multicore scalability. As technology develops according to Moore's law, obtaining optimal multicore performance will require further research in both extracting more parallelism and making sequential cores faster. Hill and Marty's model improves Amdahl's model by breaking through the limitation on speedup, but also reproduces Amdahl's argument that massive parallel computing might not be beneficial.

Acknowledgements

The authors deeply appreciate the anonymous reviewer for his insightful comments and suggestions. We would also like to thank Prof. Xian-He Sun of the Illinois Institute of Technology for his valuable talks and discussions. This research was supported partly by the National Basic Research Program of China (973) under grant 2011CB302502, and by the National Natural Science Foundation of China under grants 61003062, 60925009, 60903046 and 60921002.

References

[1] G. M. Amdahl, Validity of the single-processor approach to achieving large-scale computing capabilities, in: AFIPS Conference Proceedings, April 1967, pp. 483–485.

[2] K. Asanovic, R. Bodik, J. Demmel, et al., A view of the parallel computing landscape, Communications of the ACM 52 (10) (2009) 56–67.

[3] A. Baumann, P. Barham, P.E. Dagand, et al., The multikernel: a new OS architecture for scalable multicore systems, in: Proceedings of the 22nd ACM Symposium on Operating Systems Principles (SOSP), October 2009.

[4] M. Gebhart, B.A. Maher, et al., An evaluation of the TRIPS computer system, ACM SIGPLAN Notices – ASPLOS 2009 44 (3) (March 2009) 1–12.

[5] J.L. Gustafson, Reevaluating Amdahl's law, Communications of the ACM (May 1988) 532–533.

[6] M.D. Hill, M.R. Marty, Amdahl's law in the multicore era, IEEE Computer 41 (7) (July 2008) 33–38.

[7] X.-H. Sun, L.M. Ni, Another view on parallel speedup, in: Proc. Supercomputing'90, NY, 1990, pp. 324–333.

[8] X.-H. Sun, Y. Chen, Reevaluating Amdahl's law in the multicore era, Journal of Parallel and Distributed Computing 70 (2) (2010) 183–188.

[9] L.G. Valiant, A bridging model for multi-core computing, Journal of Computer and System Sciences 77 (1) (2011) 154–166.

[10] P.M. Wells, G.S. Sohi, Serializing instructions in system-intensive workloads: Amdahl's Law strikes again, in: IEEE 14th International Symposium on High Performance Computer Architecture, HPCA, 2008, pp. 264–275.

[11] S. Williams, A. Waterman, D.A. Patterson, Roofline: an insightful visual performance model for multicore architectures, Communications of the ACM 52 (4) (2009) 65–76.

[12] E. Yao, Y. Bao, G. Tan, M. Chen, Extending Amdahl's law in the multicore era, SIGMETRICS Performance Evaluation Review 37 (2) (2009) 24–26.

[13] http://www.cs.wisc.edu/multifacet/amdahl.