(When) Will CMPs Hit the Power Wall?

Cor Meenderinck and Ben Juurlink

Computer Engineering Department,
Delft University of Technology
Delft, the Netherlands
{cor,benj}@ce.et.tudelft.nl

Abstract. The power wall is currently one of the major obstacles computer architecture is facing. In this paper we analyze the impact of the power wall on CMP design. As a case study we model a CMP consisting of Alpha 21264 cores, scaled to future technology nodes according to the ITRS roadmap. When running at the maximum clock frequency, such a CMP would far exceed the power budget. Although power limits performance significantly, technology improvements will still provide performance growth. Amdahl's Law highly threatens this performance growth, but might not be valid for all application domains. In those cases Gustafson's Law could be valid which is much more optimistic. From our results we derive some principles to prevent CMPs from hitting the power wall.

1 Introduction

It is commonly believed that we have reached the *power wall*, meaning that uniprocessor performance improvements have come to an end due to power constraints. The main causes of the increased power consumption are higher clock frequencies and power inefficient techniques to exploit more Instruction Level Parallelism (ILP), such as wide-issue superscalar execution. Hitting the power wall is also one of the reasons why industry has shifted towards multicores or chip multiprocessors (CMPs). Because CMPs exploit explicit Thread Level Parallelism (TLP), their cores can be simpler and do not need additional hardware to extract ILP. In other words, CMPs allow exploiting parallelism in a power efficient way.

Figure 1, taken from [1], illustrates the power wall. Uniprocessors have basically reached the power wall. As argued above, multicores can postpone hitting the power wall but they are also expected to hit the power wall. Several questions arise like when will CMPs hit the power wall, what limitation will cause this to happen, what can computer architects do to avoid the problem, etc. It is generally believed that the power efficiency of CMPs can be improved by designing asymmetric or heterogeneous multicores. For example, several domain specific accelerators could be employed which are turned on and off according to the actual workload. But, is the power saving it provides worth the area cost? In this paper we try to answer those questions.

Specifically, in this paper we focus on technology improvements as they have been one of the main drivers of performance growth in the past. According to the

E. César et al. (Eds.): Euro-Par 2008 Workshops, LNCS 5415, pp. 184–193, 2009.

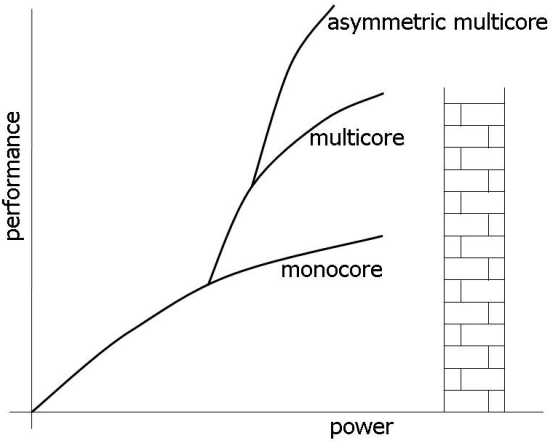

Fig. 1. The power wall problem

ITRS roadmap [2], technology improvements are expected to continue. However, power constraints might not allow us to exploit them fully. In the next section we analyze the limits of performance growth due to technology improvements with respect to power constraints. In Section 3 based on the results of our experiment we conclude that CMPs can offer significant performance improvements provided a number of principles are followed.

Of course our model is necessarily rudimentary. For example, it does not consider modifications in memory and interconnect, bandwidth constraints, nor static power dissipation due to leakage. Nevertheless, the power wall has been predicted, multicores are expected to be the remedy, asymmetric multicores have been envisioned, but to the best of our knowledge this has never been quantified. From our results we hope to derive some principles which can be the basis for future work.

2 Performance and Power of Future Multicores

To analyze the effect of technology improvements on the performance of future CMPs and to investigate the power consumption trend, the following experiment was performed. We take an Alpha 21264 chip (core and caches), scale it to future technology nodes according to the ITRS roadmap, create a hypothetical CMP consisting of the scaled cores, and derive the power numbers. Specifically, we calculate the power consumption of a CMP for full blown operation, i.e., all cores are active and run at the maximum possible frequency. Furthermore, we analyze the performance growth over time if the power consumption is restricted to the power budget allowed by packaging.

The Alpha 21264 [3] core was chosen as subject of this experiment for two reasons. First, the Alpha has been well documented in literature, providing the data required for the experiment. Second, the 21264 is a moderately sized core

lacking the aggressive ILP techniques of current high performance cores. Thus, it is a good representative of what is generally expected to be the processing element in future many-core CMPs. The relevant parameters of the 21264 core for this analysis are the following: year $t_0 = 1998$, technology node $L_{orig} = 350$ nm, supply voltage $V_{orig} = 2.2$ V, die area $A_{orig} = 314$ mm^2, dynamic power (@400 MHz) = 48 W, and dynamic power (@600 MHz) = 70 W.

2.1 Scaling of the Alpha 21264

The 21264 is scaled according to data in the 2007 edition of the International Technology Roadmap for Semiconductors (ITRS) [2]. The relevant parameters are given in Table 1. The values of the technology node and the on-chip frequency were taken from Page 79 of the Executive Summary Chapter. The on-chip frequency is based on the fundamental transistor delay, and an assumed maximum number of 12 inverter delays. The die area values were taken from the table on Page 81 of the Executive Summary. Finally, the values of the supply voltage and the gate capacitance (per micron device) were taken from the table starting at Page 11 of the Process Integration, Devices, and Structures Chapter of the roadmap.

Table 1. Technology parameters of the ITRS roadmap

	2007	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
technology (nm)	68	57	50	45	40	36	32	28
frequency (MHz)	4700	5063	5454	5875	6329	6817	7344	7911
die area (mm^2)	310	310	310	310	310	310	310	310
supply voltage (V)	1.1	1	1	1	1	0.9	0.9	0.9
$C_{g,total}$ ($F/\mu m$)	7.10E-16	8.40E-16	8.43E-16	8.08E-16	6.5E-16	6.29E-16	6.28E-16	5.59E-16
	2015	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
technology (nm)	25	22	20	18	16	14	13	11
frequency (MHz)	8522	9180	9889	10652	11475	12361	13351	14343
die area (mm^2)	310	310	310	310	310	310	310	310
supply voltage (V)	0.8	0.8	0.7	0.7	0.7	0.65	0.65	0.65
$C_{g,total}$ ($F/\mu m$)	5.25E-16	5.07E-16	4.81E-16	4.58E-16	4.1E-16	3.91E-16	3.62E-16	3.42E-16

To model the experimental CMP for future technology nodes, we scale all relevant parameters of the 21264 core. The values that are available in the ITRS, we use as such. The others we scale using the available parameters by taking the ratio between the original 21264 parameter values and the predictions of the roadmap. Below we describe in detail for each scaled parameter how this was done. The gate capacitance of the 21264 was not found in literature, thus we extrapolated the values reported in the roadmap and calculated a value of $1.1 \times 10^{-15} F/\mu m$ for 1998.

First, the area of one core was scaled. Let $L(t)$ be the process technology size for year t and let L_{orig} be the process technology size of the original core. The area of one 21264 core in year t will be $A_1(t) = A_{orig} \times \left(\frac{L(t)}{L_{orig}}\right)^2$. Figure 2 depicts the results for the time frame considered. The area of one core decreases more or

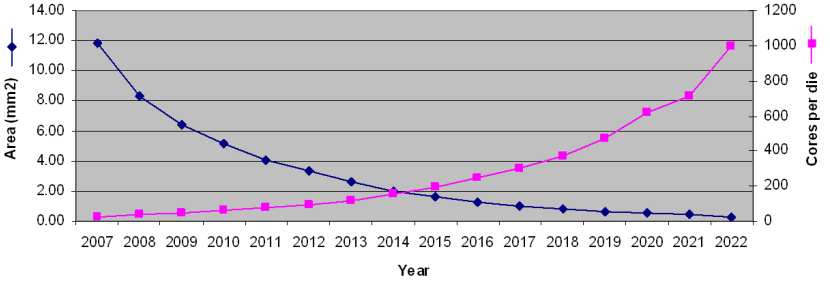

Fig. 2. Area of one core and the number of cores per die

less quadratically over time, and will be about one third of a square millimeter in 2022.

Second, using the scaled area of one core, the number of cores that could fit on a single die was calculated. The ITRS roadmap assumes a die area of $310mm^2$ for the entire time frame. Thus, the total number of cores per die in year t is $N(t) = \frac{310}{A_1(t)} = \frac{310}{A_{orig}} \times \left(\frac{L(t)}{L_{orig}}\right)^2$, which is depicted in Figure 2. The graph shows a doubling of cores roughly every three years which is in line with the expectations [4]. In 2022 our calculations predict a CMP with 999 cores.

Finally, the power of one core was scaled. Power consumption consists of a dynamic and a static part, of which the latter is dominated by leakage. The data required to scale the static power is not available to us and thus we restrict this power analysis to dynamic power. It is expected that leakage remains a problem and thus our estimations are conservative.

The dynamic power is given by $P_{dyn} = \alpha C f V^2$, where α is the transistor activity factor, C is the gate capacitance, f is the clock frequency, and V is the power supply voltage. The activity factor α of the 21264 processor is unknown and also depends on the application, but since this does not change with scaling, it can be assumed to be constant. The capacitance C (F) in the equation is different from capacitance $C_{g,total}$ $(F/\mu m)$ in Table 1, but they relate to each other as $C \propto C_{g,total} \times L$. Thus, the dynamic power at year t is calculated as:

$$P(t) = P_{orig} \times \frac{C_{g,total}(t) \times L(t)}{C_{g,total,orig} \times L_{orig}} \times \frac{f(t)}{f_{orig}} \times \left(\frac{V(t)}{V_{orig}}\right)^2. \quad (1)$$

This analysis assumes that the cores run at the maximum possible frequency. Figure 3 depicts the power of one core over time. As the curve shows it roughly decreases linearly, resulting in less than 2 W in 2022.

2.2 Power and Performance Assuming Perfect Parallelization

Now that all parameters have been scaled, it is possible to calculate the power consumption of the total CMP. It is assumed that all cores are active and thus $P_{total}(t) = N(t) \times P(t)$. Figure 3 depicts the total power over time and shows that

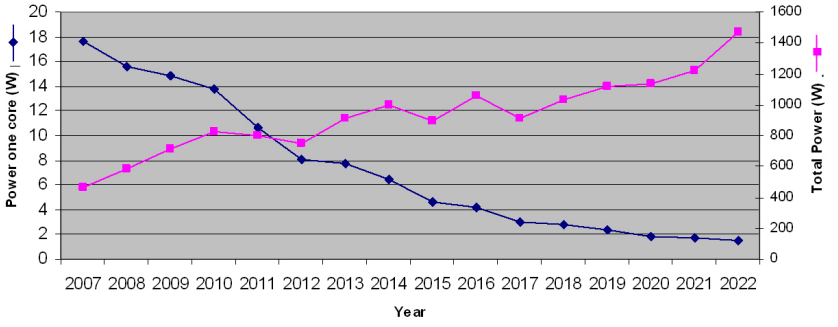

Fig. 3. Power of one core and total power of the case study CMP

the total power consumption of the modelled CMP in 2008 is 600 W, gradually increases, and reaches 1.5 kW in 2022. The roadmap predicts that the power budget allowed due to packaging constraints is 198 W. It is clear that for the entire calculated time span the power consumption of our hypothetical CMP exceeds the power budget. This is why power has become one of the main design constraints nowadays.

The figure also shows that the difference between the power budget and the power consumption of the full blown hypothetical CMP is increasing over time. That means that a large part of the technology improvement cannot be put into effect for performance growth. For example, between 2011 and 2020 technology allows doubling the on-chip frequency. However, the power consumption would increase with a factor 1.5. Thus, for equal power only a small frequency improvement would be possible.

Next we analyze the performance improvement that can be achieved by this CMP. Assuming that the application is perfectly parallelizable, the parameters that influence performance are frequency and the number of cores. In this case the speedup in year t, relative to the original 21264 core, is given by $S(t) = \frac{f(t)}{f_{orig}} \times \frac{N(t)}{1}$. This speedup is depicted in Figure 4 as the non-constrained speedup.

We are interested in the speedup achieved by CMPs that meet the power budget of 198 W. As the results show the power budget is exceeded when all cores are used concurrently at the maximum frequency. Thus, the non-constrained speedup is not achievable in practice. To meet the power budget, either the frequency could be scaled down or a number of cores could be shut down. Since both measures are linear to the speedup, the power-constrained speedup can be defined as:

$$S_{power\_constr.}(t) = \frac{f(t)}{f_{orig}} \times \frac{N(t)}{1} \times \frac{P_{budget}}{P_{total}(t)}, \tag{2}$$

where P_{budget} is the power budget allowed by packaging.

Figure 4 depicts the power-constrained performance of the case study CMP over time. To increase readability we normalized the result to 2007. The curve shows a performance growth of 27% per year. Also the non-constrained performance growth is depicted. Note that the latter is growing with 37% per year

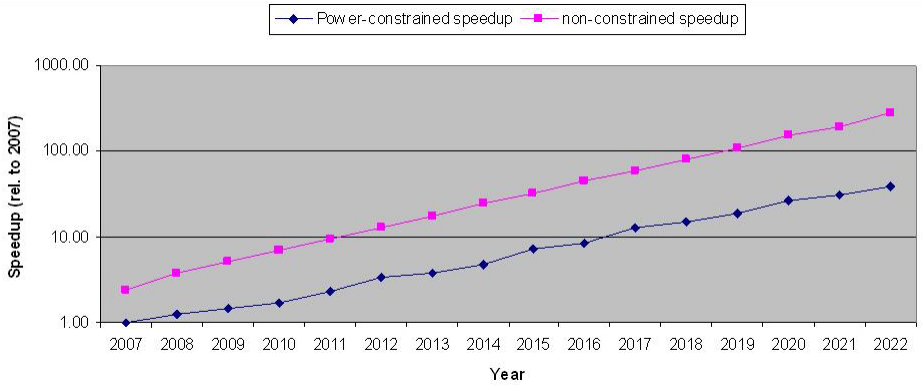

Fig. 4. Power-constrained performance growth

and that the gap between the two is increasing. The robustness of the model to variations in the input parameters is rather good. Even if all parameters we took from the ITRS roadmap vary by 20%, the predicted performance in 2022 varies between 46% and 130%.

To put these results in historical perspective we compare to Figure 2.1 of Hennessy and Patterson [5]. From the mid-1980s to 2002 the graph shows an annual performance growth of 52%. Then from 2002 the annual performance growth dropped to 20%. Considering this historical perspective, the predicted annual performance growth of 27% is a bit on the high side, but not far of. The main conclusion from these results is that although power severely limits performance, substantial performance improvements are still possible using the CMP paradigm.

2.3 Performance Assuming Non-perfect Parallelization

So far we assumed a perfectly parallelizable application. In practice this is not always the case as there might be purely serial code. If this is the case we can apply either Amdahl's Law or Gustafson's Law [6].

Both Amdahl and Gustafson proposed an equation to calculate the speedup achieved by a parallel system. At first sight they look different but Yuan Shi proved that actually they are identical [7]. However, they used different assumptions resulting in different predictions of the future. A detailed description of both laws is provided in the extended version of this paper [8].

Depending on the application domain, either Amdahl's or Gustafson's assumptions might be valid, or even something in between. First, we take Amdahl's assumptions to predict the power-constrained performance growth. We assume a symmetric CMP where all cores are being used during the parallel part. The clock frequency of all cores is equal and has been scaled down to meet the power budget. The power-constrained speedup that this symmetric CMP can achieve is given by

$$S_{Amdahl\_power-constr.\_sym.}(t) = \frac{1}{s + \frac{1-s}{N(t)}} \times \frac{f(t)}{f_{orig}} \times \frac{P_{budget}}{P_{total}(t)} \quad (3)$$

and is depicted in Figure 5. We used a serial fraction s ranging from 0.1% to 10%. The figure shows that for $s = 0.1\%$ there is a slight performance drop, compared to ideal, going up to a factor 2 for 2022. However, for $s = 1\%$ there is a performance drop of 10x for 2022 and for $s = 10\%$ there is no performance growth at all.

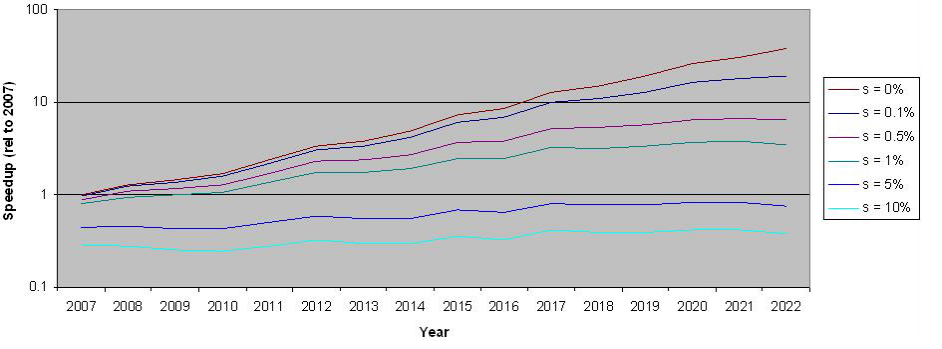

Fig. 5. Prediction of power-constrained performance growth for several fractions of serial code s assuming a symmetric CMP and with Amdahl's assumptions

Indeed we see that Amdahl's prediction is pessimistic, which is an argument for asymmetric or heterogeneous CMPs. If the serial part can be accelerated by deploying one faster core, more speedup through parallelism could be achieved. This one fast core could be an aggressive superscalar core, a domain specific accelerator, or a core that runs at a higher clock frequency than the others. For this experiment we assume identical cores but increase the clock frequency of one core during the serial part. Note that during this time the other cores are inactive and thus the power budget is not exceeded. The speedup this asymmetric CMP can achieve is given by

$$S_{Amdahl\_power-constr.\_asym.}(t) = \frac{1}{s \times \frac{f_{orig}}{f(t)} + \frac{1-s}{N(t)} \times \frac{f_{orig}}{f(t)} \times \frac{P_{total}(t)}{P_{budget}}} \quad (4)$$

and is depicted in Figure 6. Note that both this equation and Equation 3 become identical to Equation 2 if $s = 0\%$. The results show that for $s = 0.1\%$ there is only a very small performance drop compared to ideal parallelization. For $s = 1\%$ the performance drop is 2.3x while for $s = 10\%$ the performance drops 14 times compared to ideal parallelization but considerable performance growth is predicted over time. These results show that asymmetric CMPs are a good choice to improve performance, if Amdahl's assumptions are correct and if the serial fraction is larger than approximately 0.5%.

Second, we predict the power-constrained performance growth using Gustafson's assumptions. Again, we assume a symmetric CMP where all cores are being

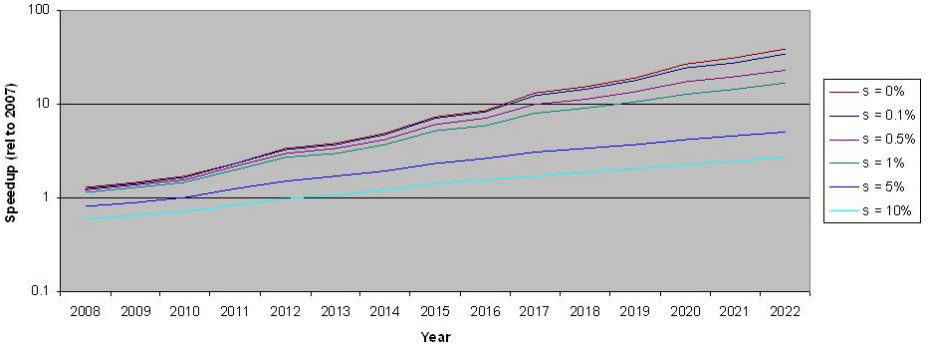

Fig. 6. Prediction of power-constrained performance growth for several fractions of serial code s assuming an asymmetric CMP and with Amdahl's assumptions

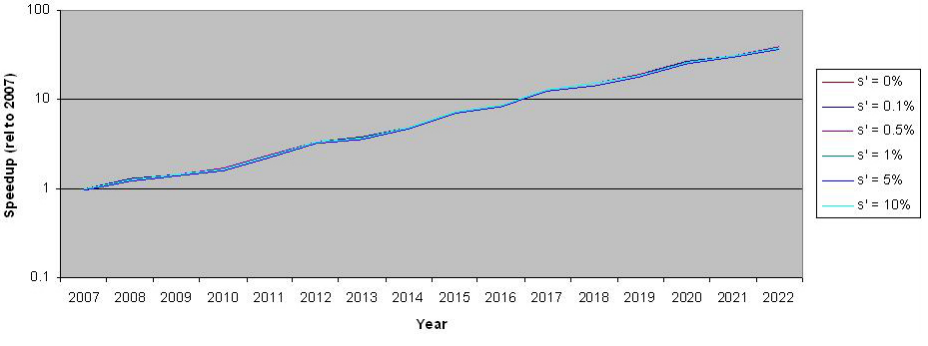

Fig. 7. Prediction of power-constrained performance growth for several fractions of serial code s' with Gustafson's assumptions

used during the parallel part. The clock frequency of all cores is equal and has been scaled down to meet the power budget. The power-constrained speedup that this symmetric CMP can achieve is given by

$$S_{Gustafson\_power-constr.\_sym.}(t) = (N + (1 - N) \times s') \times \frac{f(t)}{f_{orig}} \times \frac{P_{budget}}{P_{total}(t)} \quad (5)$$

and is depicted in Figure 7. The figure shows that for any value s' between 0% and 10% there will be no significant performance loss compared to ideal parallelization. For $s' = 10\%$ in 2022 the performance is 11% less than the ideal parallelization $s' = 0\%$ case. Thus we can conclude that symmetric CMPs are a good choice if Gustafson's assumptions are correct.

3 Conclusions

In this paper we analyzed the impact of the power wall on CMP design. Specifically, we investigated the limits to performance growth of CMPs due to

technology improvements with respect to power constraints. As a case study we modelled a CMP consisting of Alpha 21264 cores, scaled to future technology nodes according to the ITRS roadmap. It was found that in 2022 such a CMP would contain 999 cores, each consuming 1.5 W when running at the maximum possible frequency of 14 GHz. The total CMP, at full blown operation, would consume 1.5 kW while the power budget predicted by the ITRS is 198 W.

From these figures it is clear that power has become a major design constraint, and will remain a major bottleneck for performance growth. However, it does not mean that the power wall has been hit for CMPs. We calculated the power-constrained performance growth and showed that technology improvements enables a doubling of performance every three years for CMPs, which is in line with the number of cores per die.

However, there is another threat for scalable performance which is Amdahl's Law. The speedup achieved by parallelism is limited by the serial fraction s. Using Amdahl's Law we predicted the power-constrained performance growth for several fractions s. For a symmetric CMP, 1% of serial code decreases the speedup by a factor of up to 10. For an asymmetric CMP, where the serial code is executed on a core that runs at a higher clock frequency, 1% of serial code reduces the achievable speedup only by a factor of up to 2.

On the other hand, there is Gustafson's Law which uses different assumptions. The predictions with this Law are much more optimistic as even for $s' = 10\%$ in 2022 the performance loss is only 11% compared to ideal parallelization.

The question whether Amdahl's or Gustafson's assumptions are believed to be valid for a certain application domain is unresolved. Most likely some can be characterized as 'Amdahl', some as 'Gustafson' and other as something in between.

From the results of this paper we conclude that in order to avoid hitting the power wall the following two principles should be followed. First, at the architectural level power efficiency has to become the main design constraint and evaluation criterion. The transistor budget is no longer the limit but the power those transistors consume. Thus performance alone should no longer be the metric but performance per watt (or a similar power efficiency metric like performance per transistor and $BIPS^3/W$). Second, for application domains that follow Amdahl's assumptions asymmetric or heterogeneous designs are necessary. For those the need to speedup serial code remains. A challenge for computer architects is to combine speedup of serial code with power efficiency.

A CMP that follows these principles could for example look like this: a few general purpose high speed cores (e.g. aggressive superscalar), many general purpose power efficient cores (no superscalar, no out-of-order, no deep pipelines, etc.), and domain specific accelerators. The latter provides the most power efficient solution and also allows fast execution of serial code. Furthermore, dynamic voltage/frequency scaling can be applied to optimize the performance-power balance, while hardware support for thread and task management reduces the energy of the overhead introduced by parallelism. A lot more techniques and architectural directions are possible.

Summarizing, from this study we conclude that for the next decade CMPs can provide significant performance improvements without hitting the power wall, even though power severely limits performance growth. Technology improvements will provide the means, however, to achieve the possible performance improvements power efficiency should be the main design criterion at the architectural level.

Acknowledgment

The authors would like to thank Stefanos Kaxiras for his input on the methodology used in this paper.

References

1. Mendelson, A.: How Many Cores are too Many Cores? In: 3rd HiPEAC Industrial Workshop
2. ITRS: International Technology Roadmap for Semiconductors, 2007 Edition (2007), http://www.itrs.net
3. Kessler, R.: The Alpha 21264 Microprocessor. Micro 19(2), 24–36 (1999)
4. Stenström, P.: Chip-multiprocessing and Beyond. In: Proc. 12th Int. Symp. on High-Performance Computer Architecture, pp. 109–109 (2006)
5. Hennessy, J., Patterson, D.: Computer Architecture - A Quantative Approach, 4th edn., p. 3. Morgan Kaufman Publishers, San Francisco (2007)
6. Gustafson, J.: Reevaluating Amdahl's law. Communications of the ACM 31(5), 532–533 (1988)
7. Shi, Y.: Reevaluating Amdahl's Law and Gustafson's Law (1996), http://www.cis.temple.edu/~shi/docs/amdahl/amdahl.html
8. Meenderinck, C., Juurlink, B.: (When) Will CMPs hit the Power Wall? Technical report, Delft University of Technology (August 2008), http://ce.et.tudelft.nl/publications.php