Amdahl's Law in the 3-D Era

L. Yavits, A. Morad, R. Ginosar

Abstract—This work studies the effect of 3-D Chip Multiprocessor (CMP) integration on Amdahl's law. The influence of vertical communication and thermal gradients on CMP performance and scalability is studied from Amdahl's law perspective. We find that a fast vertical connectivity enabled by 3-D implementation shifts the optimum CMP configuration towards the larger number of lighter cores, thus improving CMP scalability relative to 2-D implementation. However, we also show that a high level of parallelism may lead to high peak temperatures, even in smaller scale 3-D CMPs, thus limiting 3-D CMP scalability and calling for different, non-CMP architectures.

Index Terms— Chip Multiprocessor, Multicore, Amdahl's Law, 3-D Integrated Circuits.

1 Introduction and Related Work

Power consumption and off-chip memory bandwidth are the main but not the only factors limiting the scalability of Chip Multiprocessors (CMP). On-chip inter-core communication and serial-to-parallel synchronization [13] are other significant constraints that limit the speedup of a CMP architecture as the core count grows and task parallelism is enhanced.

As integration driven by device scaling slows down [8], three-dimensional (3-D) integration arises as a natural step in CMP evolution. 3-D allows increasing transistor density by vertically integrating a number of dies with a high-speed massively parallel interface using through-silicon vias (TSV). The result is a significant reduction of interconnect both within each die and across silicon layers [2], relaxing on-chip bandwidth constraints. For instance, processing cores can be placed on multiple silicon layers to reduce the inter-core communication and serial-to-parallel synchronization latency and power. 3-D integration can also mitigate off-chip memory bandwidth restrictions by stacking one or multiple DRAM layers above CMP layers. A conceptual 3-D CMP featuring embedded multilayer 3-D DRAM is presented in Fig. 1.

Unfortunately, 3-D integration cannot eliminate the 'power wall.' Power consumption does decrease to a certain extent due to shortening the interconnect wires, but with power scaling slowing down, stacking a number of CMP layers necessarily results in a significant increase of power density. Growing power density leads to higher temperatures, which strongly affect the performance and reliability of 3-D designs. For example, placing DRAM above CMP layers might be thermally prohibitive because of hot spots where temperature may rise above the DRAM operational range (85℃-95℃ [9]), such as in 3-D DRAM cache suggested in [2].

A classical CMP architecture paradigm includes design choices such as symmetric *vs*. asymmetric CMP [15], number of cores *vs*. core size [15], cores *vs*. cache [1] [12] etc. When designing a 3-D CMP, the computer architect must address two additional questions:

1. How does vertical communication affect the number of cores and their size?
2. How do 3-D thermal gradients affect the number of cores and their size?

This paper strives to answer these questions and quantify the impact of 3-D-specific considerations on the performance and scalability of CMP.

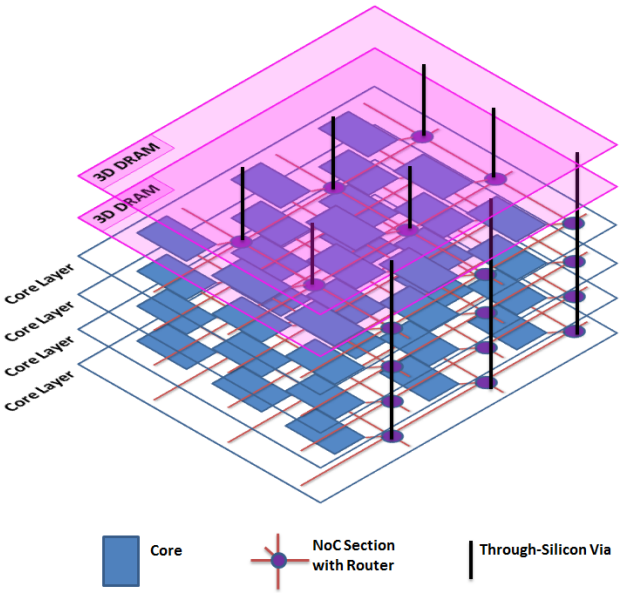

Fig. 1. Multilayer 3-D CMP with a 3-D DRAM cube stacked above it.

In recent years, there has been an extensive research on ramifications of Amdahl's law in the era of CMP. Hill and Marty [15] introduced an upper-bound analytical model for the performance and scalability of multicore and suggested an extension of Amdahl's law under a constrained area resource. Woo and Lee [3] extended the multicore performance and scalability model by addressing power consumption. Cassidy and Andreou [1] further developed the framework to account for optimal area allocation between core and memory, while Loh [7] ex-

- *Leonid Yavits (*), E-mail: yavits@tx.technion.ac.il.*
- *Amir Morad (*), E-mail: amirm@tx.technion.ac.il.*
- *Ran Ginosar (*), E-mail: ran@ee.technion.ac.il.*

(*) Authors are with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel.

tended Hill and Marty's model by adding the cost of the "uncore" resources. Chung *et al.* [4] extended the multicore corollary of Amdahl's law for heterogeneous architectures (including accelerators, such as FPGA, ASIC or GPU in addition to conventional processing cores). Eyerman and Eeckhout [16] augmented Amdahl's law by including execution of critical sections. We studied the effects of communication and synchronization on performance and scalability of a multicore [13]. Wang and Skadron [11] added supply voltage and operating frequency to Hill and Marty performance model. Recently, Ananthanarayanan *et al.* [6] extended Amdahl's law to account for process variations.

In this work, we study the effects of 3-D integration on performance and scalability of a multicore from Amdahl's law perspective. We attempt to quantify the overall multicore performance gain due to shorter inter-core distances enabled by 3-D integration. Following the methodology established by [1][3][15], we use analytical modeling to study the effect of vertical communication on the core count and size in 3-D CMP.

Our second contribution is an assessment of the limitations of CMP performance and scalability imposed by 3-D thermal gradients from Amdahl's law perspective. We employ a combination of analytical modeling and thermal simulation using the HotSpot simulator [10] to study the effects of temperature on the core count and size in 3-D CMP.

The rest of this paper is organized as follows. Section 2 analyses the effects of vertical communication. Section 3 presents the thermal analysis. Section 4 offers conclusions.

2 Effect of Vertical Communication on CMP Performance and Scaling

Hill and Marty model [15] does not account for the effects of inter-core data transfer. To assess its influence on the performance and scalability of CMP, we have suggested [13] modifying Amdahl's software model of Fig. 2(a), augmenting it to reflect the inter-core communication and serial-to-parallel synchronization as illustrated in Fig. 2(b).

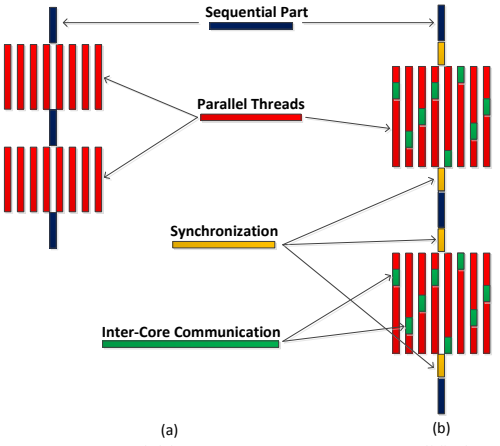

Fig. 2. (a) Amdahl's execution model (b) Modified Amdahl's execution model accounting for synchronization and inter-core communication

We propose a high-level analytical performance model that isolates and emphasizes the effects inherent to 3-D integration. Hence, similarly to Hill and Marty, we assume that the constrained n Base Core Equivalents (BCE) area resource is allocated entirely to processing cores, and no resource is spent on "uncore" components.

Hill and Marty model the performance of a single core of r-BCE size as $Perf(r) = r^{\alpha} = r^{0.5}$ [15]. A 3-D implementation may lead to faster processing cores [2][5], mainly because of shortening the inter-block distances through multi-layer partitioning of the core. We can factor in this performance improvement by adopting a power law exponent α greater than 0.5. In our model, we assume $\alpha = 2/3$.

The CMP execution time can be written as follows [13]:

$$Speedup_{sym} = \frac{r^{\alpha}}{(1-f) + \frac{f}{n_c} + \frac{f_C}{n_c} + f_S} \qquad (1)$$

where f is the parallel portion of the code, n_c is the number of cores, and f_C and f_S are the connectivity and synchronization intensities, defined as the ratio of inter-core communication and serial-to-parallel synchronization times respectively to sequential execution time [13].

Both sequential-to-parallel synchronization and inter-core communication times depend on the communication network (NoC) delay. Assuming a 2D mesh NoC, both synchronization and connectivity intensities can be presented as a function of the number of cores [7]:

$$f_C = O(\sqrt{n_c});\ f_S = O(\sqrt{n_c}) \qquad (2)$$

The speedup of a symmetric multicore as a function of core size r is presented in Fig. 3(a) (2D model) for $\alpha = 2/3$, $n = 256$ and $f = 0.99$ and 0.999. The Hill and Marty model speedup is also shown for reference.

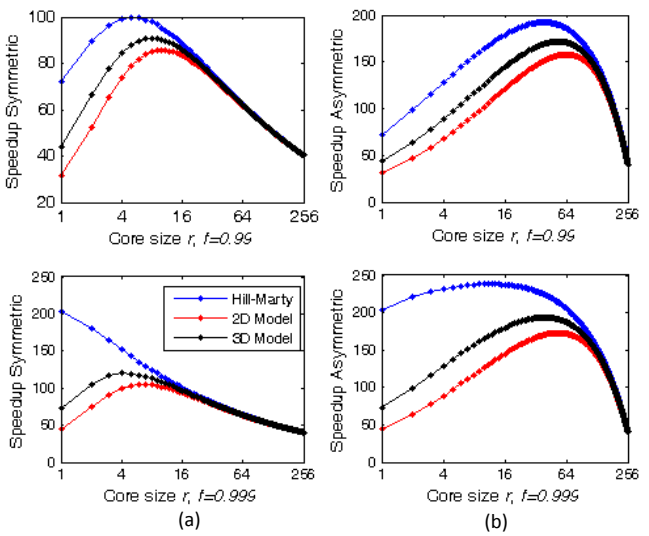

Fig. 3. The speedup of the (a) symmetric and (b) asymmetric multicore vs. core size r

Similarly, the speedup of the asymmetric multicore with one sequential core of size r and $n - r$ parallel sin-

gle-BCE cores can be written as [13]:

$$Speedup_{asym} = \frac{r^{\alpha}}{(1-f) + \frac{fr^{\alpha}}{r^{\alpha}+n-r} + \frac{f_C(n_c)}{n-r+1} + f_S(n_c)} \quad (3)$$

The speedup of a 2-D asymmetric multicore is shown in Fig. 3(b).

Sequential-to-parallel data synchronization and inter-core communication affect multicore performance in two ways. First, the overall speedup is lower than that predicted by Hill and Marty, because our model accounts for the time needed for sequential-to-parallel synchronization and inter-core communication. Second, the optimum configuration shifts from the larger number of lighter cores to a smaller number of larger cores. This happens because both the connectivity and synchronization intensities grow with the number of cores n_c ((2)).

In a 3-D design, the synchronization and connectivity intensities can be presented as a function of the number of cores and silicon layers as follows [5]:

$$f_C = O\left(\sqrt{n_c/l}\right);\ f_S = O\left(\sqrt{n_c/l}\right) \quad (4)$$

where l is the number of core silicon layers. The speedup of 3-D based CMP as a function of sequential core size r is presented in Fig. 3(a) (symmetric multicore, 3D model) and Fig. 3(b) (asymmetric multicore, 3D model) for $l = 4$ silicon layers.

Consequently, we conclude that a 3-D integration improves the performance of CMP due to shorter wire distances. It also shifts the optimum configuration towards the upper-bound Hill and Marty model: when the sequential-to-parallel synchronization and inter-core communication effects are taken into account, the best speedup is achieved with a larger number of smaller cores compared to a 2-D CMP.

Result 1: Sequential-to-parallel data synchronization and inter-core communication, both inherent to parallel execution, reduce the multicore speedup and move the optimum configuration towards a smaller number of larger cores. 3-D integration improves performance and enhances scalability due to faster vertical connectivity.

3 Effect of Temperature on CMP Performance and Scaling

In this section, we present the thermal analysis of a 3-D CMP using the HotSpot simulator [10]. The inputs to HotSpot are the multicore floorplan and its power trace.

We consider a symmetric multicore with the BCE budget of 256. To create the power trace, we use a methodology based on [3] and [4]. Let P be the dynamic power consumption of a "fully blown" $R = 256$ BCE processing core. The fraction of power consumed in idle state is k $(0 \le k \le 1)$. The power of a smaller core relative to the power of the "fully blown" one is w_c $(0 \le w_c \le 1)$. Following [4], we scale w_c as a power law of the core size r:

$$w_c = \left(\frac{r}{R}\right)^{\frac{1.75}{2}} \quad (5)$$

The fraction of the smaller core's idle power normalized to the same core's overall power consumption is k_c $(0 \le k_c \le 1)$.

During serial execution, only one (the serial) core of this multicore is active, while the rest are idle. Hence the power consumption of all cores and the execution time of the serial core are as follows:

$$p_{1,1} = w_1 P;\ p_{j,i} = w_1 k_1 P;\quad T = T_1(1-f) \quad (6)$$

where T_1 is the sequential execution time on a single 1BCE core; 1,1 is the sequential core; $i, j \in 1\ to\ 16$ except for $i = j = 1$.

During the parallel execution, all cores are active. The power consumption and execution time of the individual cores hence are as follows:

$$p_{j,i} = w_1 P;\quad T = {fT_1}/{256};\quad i, j \in 1\ to\ 16 \quad (7)$$

Note that in (6) and (7), we assume for simplicity that no power is consumed during sequential-to-parallel data synchronization and inter-core communication. This assumption benefits the large scale CMPs (large number of small cores). Accounting for the interconnection network power, which can be quite significant, would have further increased the per-core power (5) of a large-scale CMP relative to a small-scale one (small number of large cores).

For the thermal analysis, we assume a four-layer CMP stack. We use Intel's Core™ 45nm processor as a reference "fully blown" core with $P = 15W$ and $R = 23mm^2$ [8]. We then consider a symmetric multicore, optimally-configured for $f = 0.999$ in accordance with Hill and Marty model, comprising 256 cores, each of size $r = {23mm^2}/{256}$ in each one of the four silicon layers. The thermal map of the upper layer is presented in Fig. 4(a). The peak temperature of this layer is around 109℃. This temperature is above the maximum operating temperature of most DRAMs (85℃-95℃), making integration of 3-D DRAM above the processor stack (as presented in Fig. 1) thermally infeasible.

Fig. 4(b) presents the peak temperature of the upper processor layer of a 3-D CMP, obtained by HotSpot simulation as a function of core size for different values of f. The figure also shows the maximum DRAM temperature. For low parallelism tasks (low thread count), the peak temperature of the upper processor layer is within DRAM operational range, since most cores are idle. However, for highly parallel tasks ($f \ge 0.999$), when most cores are active most of the time, the peak temperature of the upper processor layer is above the DRAM operational range even for a smaller-scale CMP (8 cores and above). A 3-D implementation where DRAM cannot be placed atop a multi-layer CMP fails one of its essential purposes, which

is breaching the memory and bandwidth walls.

Result 2: Peak temperatures of 3-D CMP grow with parallelism and the number of cores. Inherent 3-D thermal constraints also limit the CMP scalability, restricting the practical CMP configuration to a smaller number of larger cores. CMPs with a very large number of small active cores, targeted for highly parallelizable applications, are less suitable for 3-D implementation.

Implication: Increasing parallelism (as suggested by Hill and Marty in [15]) in 3-D CMP without addressing its thermal outcome has in fact an adverse effect on its scalability and performance. Hence, multicore designers should seek ways to reduce the peak temperatures and hotspots, to enable an efficient 3-D integration of a large scale CMPs. For instance, a radically different, non-CMP 3-D architecture employing associative processing has been shown to deliver high performance while maintaining temperatures compatible with DRAM layers [14].

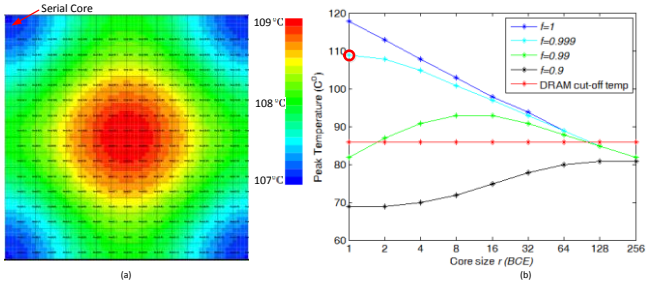

Fig. 4 (a) Thermal map of the optimally-configured (r = 1BCE) CMP, f = 0.999; (b) Peak Temperature in the upper silicon layer vs. core size r for different f; the peak temperature of Fig. 4(a) configuration is circled in red.

4 Conclusions

As integration driven by device scaling slows down and the memory wall looms, 3-D integration becomes a natural step in CMP evolution. This work studies the effect of 3-D multicore implementation on its performance and scalability from the perspective of Amdahl's law. We focus on two aspects of 3-D implementation: shortening wire distances through vertical co-location of the critical components in separate silicon layers and enabling vertical connectivity, and thermal limitations of a multilayer 3-D CMP and DRAM integration.

We find that 3-D implementation of CMP leads to shifting the optimum CMP configuration towards a larger number of smaller cores, thus improving 3-D CMP scalability and performance relative to conventional 2-D CMP. This improvement is enabled by a faster vertical connectivity.

We further find that the peak temperatures of 3-D CMP grow with the number of cores and with parallelism. Hence, 3-D CMP scalability is limited by thermal gradients, pushing the practical CMP configuration towards a smaller number of larger cores. CMPs with a large number of small active cores, targeted for highly parallelizable applications, are less suitable for 3-D implementation

An implication of our research is that multicore designers should seek ways to reduce the peak temperatures and hotspots, to enable efficient 3-D integration of larger scale CMP. Increasing parallelism (as suggested by Hill and Marty in [15]) in 3-D CMP without addressing its thermal outcome has an adverse effect on CMP scalability and performance. Other architectures, such as [14], may prove more suitable for highly parallel 3-D implementations.

Acknowledgment

This research was partially funded by the Intel Collaborative Research Institute for Computational Intelligence and by Hasso-Plattner-Institut.

References

[1] A. Cassidy and A. Andreou, "Beyond Amdahl Law - An objective function that links performance gains to delay and energy", IEEE Transactions on Computers, vol. 61, no. 8, pp. 1110-1126, Aug 2012.

[2] B. Black *et al.* "Die stacking (3-D) microarchitecture", MICRO-39, 2006.

[3] D. H. Woo and H. H. Lee. "Extending Amdahl's law for energy-efficient computing in the many-core era." Computer 41.12 (2008): 24-31.

[4] E. Chung *et al.* "Single-chip heterogeneous computing: Does the future include custom logic, FPGAs, and GPGPUs?", 43rd Annual IEEE/ACM International Symposium on Microarchitecture, 2010.

[5] F. Li *et al.*, "Design and management of 3-D chip multiprocessors using network-in-memory", ACM SIGARCH Computer Architecture News 34.2 (2006): 130-141.

[6] G. Ananthanarayanan *et al.* "Amdahl's Law in the Era of Process Variation", Int. Journal of High Performance Systems Architecture, 2013

[7] G. Loh, "The cost of uncore in throughput-oriented many-core processors", Workshop on Architectures and Languages for Throughput Applications. 2008.

[8] H. Esmaeilzadeh, *et al.* "Power challenges may end the multicore era." Communications of the ACM 56.2 (2013): 93-102.

[9] http://www.micron.com/products/dram

[10] K. Skadron et al., "Temperature-aware microarchitecture." APM SIGARCH Computer Architecture News. Vol. 31. No. 2. APM, 2003.

[11] L. Wang and K. Skadron, "Dark vs. Dim Silicon and Near-Threshold Computing Extended Results.", http://www.cs.virginia.edu/~lw2aw/files/Wang_TR_2013-01.pdf

[12] L. Yavits, A. Morad, R. Ginosar, "Cache Hierarchy Optimization," IEEE Computer Architecture Letters, 2013

[13] L. Yavits, A. Morad, R. Ginosar, "The effect of communication and synchronization on Amdahl's law in multicore systems", Parallel Computing journal, 2013.

[14] L. Yavits, A. Morad, R. Ginosar, E. Friedman, "Associative Processor Thermally Enables 3-D Integration of Processing and Memory", http://webee.technion.ac.il/~ran/papers/YavitsThermalAssociative.pdf

[15] M. Hill, M. Marty. "Amdahl's law in the multicore era", Computer 41.7 (2008): 33-38.

[16] S. Eyerman and L. Eeckhout. "Modeling critical sections in Amdahl's law and its implications for multicore design." ACM SIGARCH Computer Architecture News. Vol. 38. No. 3. ACM, 2010.