Timing, Energy, and Thermal Performance of Three-Dimensional Integrated Circuits

Shamik Das, Anantha Chandrakasan, and Rafael Reif
Microsystems Technology Laboratories
Massachusetts Institute of Technology
60 Vassar St. Rm. 39-625
Cambridge, MA 02139
shamikd,anantha,reif@mtl.mit.edu

ABSTRACT

We examine the performance of custom circuits in an emerging technology known as three-dimensional integration. By combining multiple device layers with a high-density inter-layer interconnect, 3-D integration of a given circuit is expected to provide better timing and energy performance relative to a single-wafer implementation of the same circuit. In this paper, we show that by using our performance-driven design tool for 3-D ICs, the interconnect energy dissipation of standard-cell circuits can be reduced by 24% to 42% using two to five device layers respectively. Similarly, the interconnect energy-delay product can be reduced by 30% to 50%.

At the same time, thermal performance in 3-D ICs is expected to be a critical issue. By incorporating thermal management and analysis into our placement tool, we may investigate the thermal scalability of 3-D integration. We find that the thermal performance actually can be improved with the use of a modest number of additional device layers. Also, we show that the absolute die temperature can be controlled through the use of extra silicon.

Categories and Subject Descriptors

B.7.1 [**Integrated Circuits**]: Types and Design Styles—*VLSI*, *advanced technologies*; B.7.2 [**Integrated Circuits**]: Design Aids—*placement and routing*

Keywords

3-D integration, 3-D IC, timing, energy, thermal optimization

1. INTRODUCTION

Timing, energy consumption, and area are the main parameters of interest for any digital circuit designer. Circuit timing is usually considered foremost; most often, the timing requirement is stated as a maximum cycle time or minimum operating frequency. However, at the placement stage of the design process, one commonly targets the best possible timing performance regardless of the constraint – it is important to be able to guarantee that the timing specification is met. Energy optimization is then performed secondarily. Thermal characteristics are typically managed rather than optimized.

GLSVLSI'04, April 26–28, 2004, Boston, Massachusetts, USA.

By contrast, sometimes it can be desirable to minimize energy consumption while merely satisfying the timing constraint. For example, a designer of a wireless cryptographic card may care most about energy per encryption. The real-time requirement may be lax enough to permit energy optimization solely, or there may be a significant timing constraint. Similarly, there may be reason to improve thermal performance beyond what is required to meet digital specifications, especially where mixed-signal circuits may be involved.

The demand for timing-constrained energy optimization is not of recent origin; however, as CMOS technology improves, the nature of energy consumption in digital ICs changes. Since traditional CMOS interconnect performance does not scale as well as device performance, new technologies and computer-aided design solutions must focus on interconnect [1]. Specifically, as new interconnect technologies appear, design tools must be able to exploit them for optimal performance.

In this paper, we present a performance optimization tool for a promising new technology: **three-dimensional integration**. A 3-D IC is one that is designed using more than one wafer or active device layer. These device layers are stacked so that transistors may be wired not only to other transistors in the same wafer plane, but also to transistors in adjacent planes. A circuit that is placed and routed in multiple wafers will have a wire-length distribution that is shifted towards the local wires when compared with the same circuit placed and routed on a single wafer [2]. With wires that are generally shorter in 3-D ICs, both switching energy and cycle time are expected to be reduced. It would, of course, be desirable to trade off cycle-time reduction for decreased energy consumption. A circuit design methodology that fixes the cycle time while optimizing energy consumption or thermal profile allows us to evaluate properly the energy characteristics of 3-D ICs.

Our design tool can optimize the energy consumption and the thermal profile of a standard-cell circuit layout under a supplied timing constraint. We focus on the interconnect-related components of energy consumption that can be affected by placement-based optimization. For each of a pair of circuits, we obtain a placement using one to five wafers. We compare the energy consumption profile under our timing-constrained approach with that of the same circuit under a timing-optimized approach and show that an appreciable decrease in interconnect energy consumption can be obtained. Furthermore, this improvement scales as more wafers are used in the 3-D IC.

We also examine the thermal characteristics of 3-D ICs. For a given circuit, we perform placement-based optimization of the thermal profile of the circuit. We show that thermal metrics such as maximum die temperature can actually be improved using multiple wafers, under certain conditions. We quantify the extent to which design methodology can control the thermal behavior of 3-D ICs,

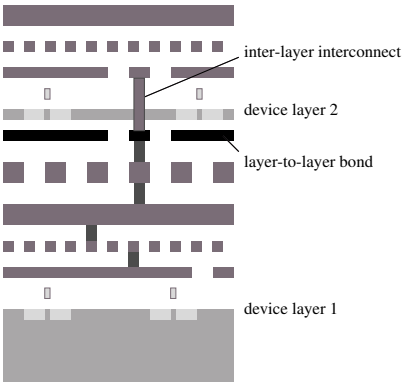

Figure 1: Wafer-bonded structure with two device layers and copper interconnect interface. (Figure courtesy A. Fan, MIT.)

and thus the extent to which technological measures must be implemented in order to make 3-D integration a workable technology.

2. BACKGROUND ON 3-D INTEGRATION

Three-dimensional integration comprises a class of fabrication technologies, each of which aims to overcome the planar limitation of conventional single-wafer ICs. Specifically, on a silicon substrate, any given transistor has a limited number of "nearest neighbors," those transistors to which it can connect with a minimum-length wire. By stacking individual transistors or whole substrates on top of one another, one may increase the number of nearest neighbors, provided that one has a suitable means for interconnecting transistors on different levels of the stack. Because of the greater number of nearest neighbors available to individual transistors, a circuit fabricated as a 3-D IC may have more short wires and fewer long wires compared to the same circuit fabricated as a conventional IC [2].

Several technologies have been proposed for 3-D integration [3, 4, 5, 6, 7]. Of these, approaches such as wafer bonding [6, 7] offer the most promising inter-wafer interconnect in terms of circuit performance [8] due to the high density and relatively low parasitic values associated with the interconnect. A diagram of a two-wafer bonded IC is shown in Figure 1. In this bonding method, individual wafers are fabricated by conventional means; the inter-wafer interconnect is formed by patterning vias that connect to the top-level metal of the bottom wafer and the first-level metal of the top wafer. Copper bonding pads are then patterned for electrical and mechanical connectivity between the wafers. The two wafers are then bonded under heat and pressure.

For the purposes of this paper, we assume a 0.18 μm technology. The individual wafers are fabricated as SOI wafers with a thickness of one to two μm. The pitch of the inter-wafer interconnect is two μm, as dictated by the alignment tolerance of the bonding process. Such dimensions are achievable with modern fabrication technologies [6, 7].

3. PLACEMENT-BASED METHODOLOGY

At current technology nodes, switched capacitance dominates the energy consumption of digital ICs. Furthermore, this capacitance increasingly comes from wires. Since 3-D integration achieves a fundamental shift in the distribution of wire lengths, an energy strategy that focuses on minimizing switched capacitance will be useful for evaluating 3-D ICs. Placement is a natural stage at which to perform this sort of wire-length optimization.

Additionally, the thermal profile of a circuit is dictated by its energy profile and its packaging. Placement-based manipulation of the energy profile, including the energy associated with switched capacitance, will allow us to examine the trade-offs between thermal optimization and energy optimization in 3-D ICs.

3.1 The Placement Tool

We have previously developed a placement and routing tool for 2-D (i.e. conventional) and 3-D integrated circuits [9]. Called **PR3D**, this tool is capable of targeting standard-cell designs for a single wafer or for multiple wafers. PR3D, as a conventional wire-length-driven placement tool [10], is competitive with modern industry and academic tools such as Dragon [11], Capo [12], and Cadence Silicon Ensemble. As a design tool for 3-D ICs, PR3D is capable of wire-length-driven placement onto a user-specifiable number of wafers with inter-wafer interconnect parasitics that are also user-specified.

The core placement algorithm is refinement by recursive bisection of the net list. Specifically, the circuit net list is represented by a hypergraph, with standard cells becoming nodes and wires becoming hyperedges. The die area is partitioned recursively into halves such that the number of nets crossing any partition is minimized [13].

In order to optimize the energy performance, we have extended the placement algorithm to include switching activity. Specifically, the energy consumption of a net i is given by

$$E_i = N_i \left(C_{is} + \sum_{j \neq i} M_{ij} C_{ij} \right) V_{DD}^2,$$

where N_i is the number of 0-to-1 transitions, C_{is} is the capacitance of the net to the substrate, C_{ij} is the coupling capacitance to net j, M_{ij} is a Miller factor that accounts for signal correlations between nets i and j, and V_{DD} is the supply voltage. The switching activity is given by the average number of transitions per unit time or per cycle.

Since the capacitance C_{is} essentially follows the net length, the energy consumption may be reduced by weighting each net according to its activity. We extend our placement tool to minimize the weighted sum of the nets crossing a partition. Thus, nets with high activity are less likely to be cut by a partition. This leads to high-activity nets being highly localized and therefore shorter and less capacitive. (The coupling capacitance C_{ij}, while important in computing energy consumption, is difficult to determine before routing is complete. However, we assume that it too will be reduced if we reduce the lengths of highly-active wires.)

At the same time, we have also extended PR3D to manage timing performance during placement. We utilize a combination of net-based and path-based approaches such as in [14, 15]. Separate approaches are employed for timing optimization and for timing constraint.

For timing optimization, we use a standard path-based counting technique. That is, we seek to minimize both net cut and path cut during recursive bisection. Nets are weighted according to the number of critical paths on which they lie. Furthermore, if a given path exceeds a fixed number of path cuts, the nets on that path are prohibited from being cut further.

Conversely, for timing constrained optimization of energy or temperature, delay is not part of the cost function. We therefore seek to minimize net cut, weighted as before (i.e. either unweighted net-cut or net-cut weighted by switching activity); however, we insert a timing-analysis step between partitionings, and if any critical path exceeds 95% of its allotted delay, the nets on that path are also prohibited from further cuts.

For thermal optimization, we extend the methodology of [16] in order to optimize 3-D IC placements. Specifically, energy consumption at any given physical location in a circuit translates into a rise in temperature at that location as the energy is dissipated into the substrate as heat. The temperature distribution within any ma-

terial component of a chip may be computed using the steady-state heat diffusion equation

$$k \cdot \nabla^2 T + g(x, y, z) = 0,$$

where T is the temperature distribution, g is the power density distribution, and k is the thermal conductivity of the material. This equation may be solved by using the finite-difference method and discretizing the 3-D IC into an m-by-m-by-p grid of $n = m^2 p$ nodes (we take $m = 50$ for lateral temperature resolution, and p equal to the total number of distinct material layers over all wafers, with extra layers allocated for bulk materials such as the bottom substrate). The result is a matrix equation $GT = P$, where G is an n-by-n matrix of thermal conductances connecting adjacent nodes, T is the temperature at each of the nodes, and P is the power dissipation at each node.

Given a circuit layout and operating frequency, the power dissipation P_k is known, and the temperature $T_k = G \backslash P_k$ may be computed (in our case, using the preconditioned conjugate gradient method). More importantly, given a desired thermal distribution T_d, a power constraint $P_d = GT_d$ may be computed. Placement optimization of 2-D ICs using this power constraint is carried out in [16].

For 3-D ICs, we assume a conventional package where the bottom substrate is attached to a heat spreader and heat sink. Numbering the wafers consecutively from 1 to n with wafer 1 adjacent to the sink, the average temperature of wafer i must exceed that of wafer $i-1$, because the heat from the ith wafer must flow through wafers $i-1$ through 1 before being dissipated into the sink. Therefore, if the distribution T_d is desired to be uniform over all wafers, the resulting power constraint is zero over wafers 2 through n. So instead of attempting to obtain a uniform thermal distribution for the entire circuit, we focus on the within-wafer variation for each wafer. To manage wafer-to-wafer thermal gradients, we attempt to place most of the energy dissipation close to the heat sink. Specifically, when partitioning a sub-circuit placement into wafers i and $i+1$, we minimize the energy consumption on wafer $i+1$, subject to the constraint that equal areas of standard cells are placed on each wafer.

4. CASE STUDIES

In order to evaluate the effectiveness of our energy and thermal optimization methodologies, we placed and routed two circuits. For each circuit, we obtained three layouts, the first optimized for minimum cycle time, the second optimized for minimum switching energy under a timing constraint, and the third optimized for best thermal profile under the same timing constraint. The circuits are supplied in Verilog format and compiled to cells using Synopsys Design Compiler. During this synthesis, Design Compiler is supplied with the timing constraint that is later used for energy optimization by PR3D.

Design Compiler is also used to assess the activity factors of the nets in the design, by using a number of representative test inputs in gate-level simulation. The activity factors are produced in SAIF format and imported into PR3D. While this method of obtaining activity factors is not exhaustive or rigorous, the problem of rigorously determining activity factors is orthogonal to our study. Furthermore, we can expect to capture a reasonable portrait of the energy consumption profile using this technique.

Once layout is generated, extraction is performed on the layout, and the resulting transistor-level net list is simulated using Synopsys NanoSim.

4.1 2-D ICs

Figure 2 shows the energy consumption of the first circuit, a 32-bit fast Fourier transform (FFT) datapath. In the left-hand graph, switched-capacitance energy dissipation accounts for about 87% of total energy consumption, with the remainder being cell internal energy. The switching energy consists of two parts. An estimated 43% is due to switching at the cell inputs and outputs (i.e. gate and source/drain capacitances). The remaining 44% is due to wires. This layout is optimized for cycle time.

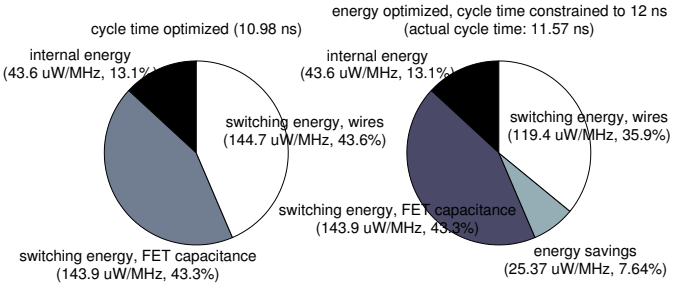

Figure 2: Energy consumption of an FFT datapath in timing-optimized vs. timing-constrained placement.

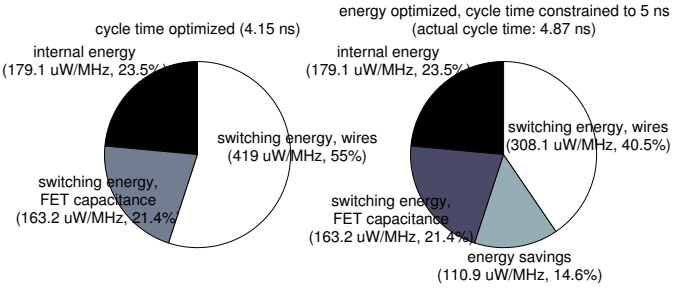

Figure 3: Energy consumption of a DES chip in timing-optimized vs. timing-constrained placement.

The right-hand graph shows the same circuit, where in this case the cycle time is constrained to 12 ns, and energy is optimized by the placement tool. While the cycle time is approximately 0.6 ns slower than in the first case, it still meets the constraint. Furthermore, the wire component of energy dissipation is reduced by 18%, leading to an overall reduction of 8%.

Figure 3 shows the energy consumption of the second circuit, a cryptographic chip implementing the Data Encryption Standard (DES). In this case, 76% of the total energy dissipation of the timing-optimized layout is due to switched capacitance, as seen in the left-hand graph. This 76% consists of 21% cell I/O switching energy and 55% wire switching energy. For this circuit, we see in the right-hand graph that while the cycle time has increased by about 0.7 ns (still meeting the constraint), the interconnect energy dissipation has been reduced by 26%, leading to an overall reduction in energy consumption by 15%.

4.2 3-D ICs

4.2.1 Energy Performance

Previous work on the emerging technology of three-dimensional integration has focused mainly on its effects on the total wire length of circuits [2, 8]. It has been shown that with a favorable technology such as wafer bonding, aggregate wire length can be reduced by 30%-50% by using two to five wafers to fabricate a given circuit [8].

However, precisely how this translates into more relevant metrics such as cycle time and energy dissipation has been unknown until now. With the use of our performance-driven design tool for 3-D ICs, we are able to characterize sample circuits with respect to these metrics.

Figure 4 shows how the interconnect energy dissipation of the

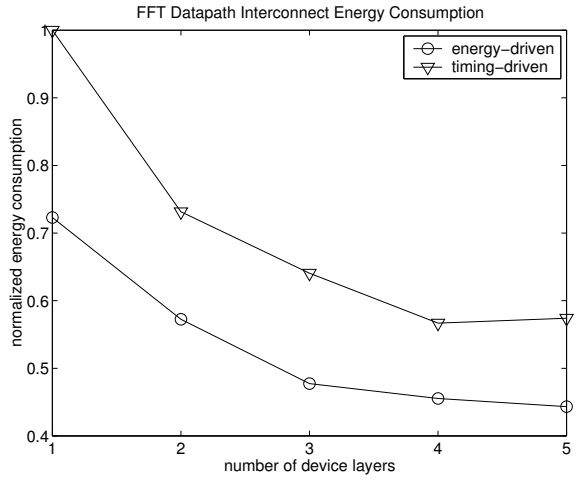

Figure 4: Interconnect energy consumption of the FFT datapath vs. number of wafers used for placement.

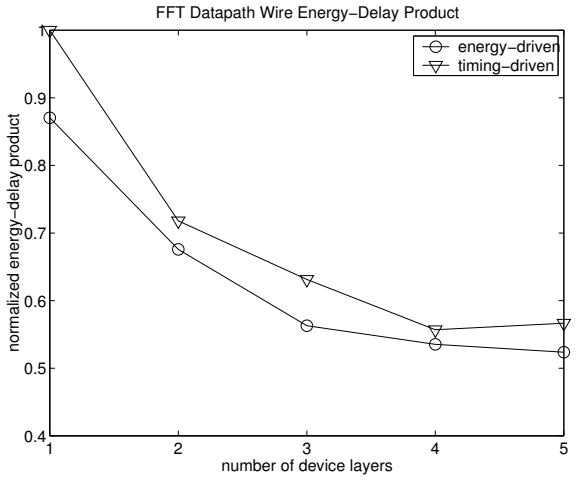

Figure 6: Interconnect energy-delay product for the FFT datapath vs. number of wafers used for placement.

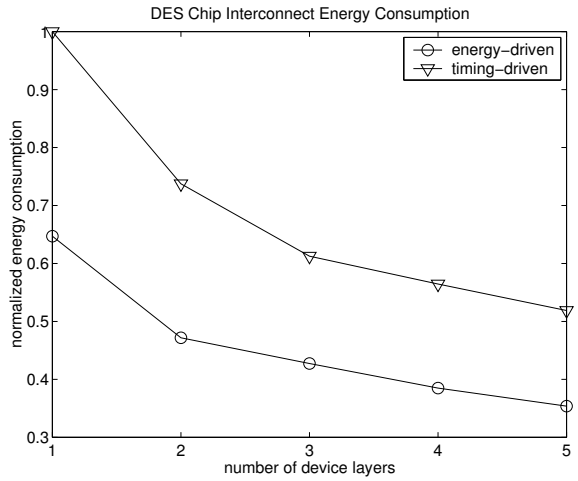

Figure 5: Interconnect energy consumption of the DES chip vs. number of wafers used for placement.

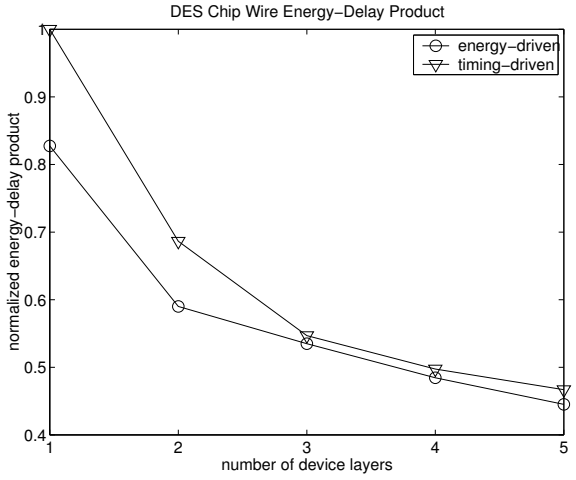

Figure 7: Interconnect energy-delay product for the DES chip vs. number of wafers used for placement.

FFT datapath circuit scales with the number of wafers. Two cases are shown: timing-driven mode and timing-constrained, energy-driven mode. In the latter mode, the FFT datapath cycle time is constrained to 12 ns. We observe that in addition to the energy savings relative to timing-driven mode, in timing-constrained mode we are able to reduce interconnect energy consumption 21% to 39% using two to five wafers respectively. Relative to a single-wafer timing-optimized design, we can reduce interconnect energy consumption by 49% by doing timing-constrained energy optimization and using five wafers.

In Figure 5, we observe similarly that for the DES chip, 27% to 45% of the interconnect energy consumption of a 2-D layout can be eliminated by targeting two to five wafers respectively. Here, relative to a single-wafer timing-optimized design, we can reduce interconnect energy consumption by 60% by doing timing-constrained energy optimization and using five wafers.

We can also measure the wire energy-delay product for both circuits, shown in Figures 6 and 7. It is clear that timing-constrained energy optimization is a win for both 2-D and 3-D integrated circuits. We can see from the graphs, however, that the cycle time of the timing-optimized versions of the two circuits is slightly better than that of the timing-constrained versions, and that this timing performance improves as more wafers are used. This is especially true for the DES chip; however, an overall penalty is incurred in the form of the extra energy consumption required to obtain optimal timing performance.

4.2.2 Thermal Performance

Figures 8 and 9 illustrate the mechanism of our thermal optimization. Figure 8 shows the temperature of the uppermost die of a three-wafer FFT placement. In the energy-optimized case, there is a hot spot that arises from shortening all the highly-active wires. In Figure 9, the origin of the hot spot is clear from the energy distribution. Thermal optimization spreads the energy consumption over the entire die, so that the hot spot is reduced or eliminated. We have assumed a conventional package with a heat sink extraction capability of $10^{-4}\mathrm{m}^2\mathrm{K/W}$, achievable with currently-available technology [17]. In all analyses, the circuit is run at 80 MHz in an ambient temperature of $25^{\circ}\mathrm{C}$.

Figures 10-15 show the thermal performance of the FFT datapath using one to five wafers. In the first set of figures, we assume that the overall footprint of the die is unchanged as we scale the number of wafers (as may be the case in an I/O-limited situation). Figure 10 shows the temperature of each die for both placements. In Figure 11, we plot the absolute temperature difference (maximum temperature minus minimum temperature over the entire circuit)

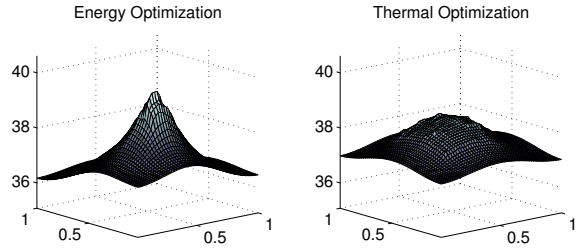

Figure 8: Celsius die temperature of the top wafer of a three-wafer placement of the FFT datapath.

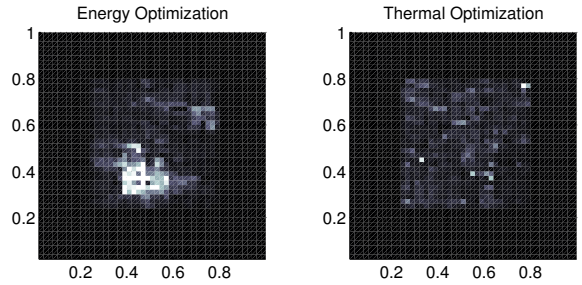

Figure 9: Energy distribution of the top wafer of a three-wafer placement of the FFT datapath.

against the number of wafers used. Figure 12 gives the wafer-to-wafer average temperature difference (i.e. average temperature of the hottest wafer minus average temperature of the coolest wafer). Figures 13-15 provide the same data for the case where the overall footprint of the die scales inversely with the number of wafers used, as may be expected for general-purpose 3-D ICs.

In both cases, we see that there is a trade-off between best energy performance and best thermal performance (which for now we construe to mean "smoothest" thermal distribution). Specifically, we see that the mean temperature of the thermally-optimal case is higher than the case where energy is optimized. This is because in order to distribute the energy consumption uniformly, some highly-active wires must be made longer, thus increasing energy consumption. For this circuit in particular, there is about 60% additional interconnect energy dissipation in the thermally-optimal case. Also, it can be seen from the graphs that the improvement in thermal performance obtained by doing thermal optimization for 3-D diminishes as more wafers are used, and asymptotic limits are reached. Thus, the design choice of whether to go for energy optimization or thermal optimization is dictated by whether a smooth thermal profile is required (as may be the case for mixed-signal circuits or digital circuits with a severe hot-spot problem) versus whether a lower mean temperature is desired.

Furthermore, we do see by comparing the fixed-die and scaled-die cases that it is possible to control the die temperature by using extra silicon. Since the energy consumption improves as the number of wafers is increased, the die area can be scaled in a proportional fashion so as to maintain a constant average temperature. However, if it is desired not to sacrifice silicon for thermal purposes, the runaway thermal behavior of Figure 13 must be controlled by advanced packaging and cooling techniques.

5. CONCLUSION

We have presented a performance-driven design tool that allows exploration of the three-dimensional integrated circuit design space. We find that by exploiting 3-D integration, we can reduce interconnect energy consumption by 24% to 42% using two to five wafers. In combination with the energy optimization for 2-D designs, we find that we can eliminate as much as 60% of the wire energy dissipation of a single-wafer timing-optimized circuit by using 3-D integration.

The thermal outlook for 3-D ICs is less clear. We have quantified the trade-off that exists between mean circuit temperature and smoothness of the temperature profile. Furthermore, we have seen that the benefit of thermally optimizing a 3-D IC placement converges toward the thermal performance of an energy-optimized 3-D IC as more wafers are used. However, we have also quantified how the runaway thermal behavior expected of 3-D ICs can be controlled by the use of extra silicon.

The conclusion is that 3-D integration promises significant energy savings that are physically realizable without any significant show-stoppers. Further research must be done to provide optimal manufacturing, packaging, and yield solutions in order to bring this technology to the designer and to the commercial market.

6. ACKNOWLEDGMENTS

This work has been carried out as part of the IFC Research Program at MIT, and is supported in part by MARCO, its participating companies, and DARPA under contract 2003-IT-674 and grant BX-8771. The authors would also like to thank the conference reviewers, as well as Brian Ginsburg, Nisha Checka, Andy Fan, Tan Mau Wu, and Anne French, for their helpful comments on this paper.

7. REFERENCES

[1] M. T. Bohr. Interconnect scaling – the real limiter to high-performance ULSI. In *Proc. IEDM*, pages 241–244, 1995.

[2] A. Rahman and R. Reif. System-level performance evaluation of three-dimensional integrated circuits. *IEEE TVLSI*, 8(6), 2000.

[3] C. W. Eichelberger. Three-dimensional multichip module system. United States Patent 5,111,278, May 1992.

[4] G. Roos *et al.* Manufacturability of 3D-epitaxial-lateral-overgrowth CMOS circuits with three stacked channels. *Microelectronic Engineering*, 15:191–194, 1991.

[5] V. Subramanian *et al.* Controlled two-step solid-phase crystallization for high-performance polysilicon TFTs. *IEEE Electron Device Letters*, 18:378–381, Aug. 1997.

[6] A. Fan, A. Rahman, and R. Reif. Copper wafer bonding. *Electrochemical and Solid-State Letters*, 2:534–536, 1999.

[7] Y. Kwon *et al.* Dielectric glue wafer bonding for 3D ICs. In *Proc. MRS*, Spring 2003.

[8] S. Das, A. Chandrakasan, and R. Reif. Three-dimensional integrated circuits: Performance, design methodology, and CAD tools. In *Proc. ISVLSI*, 2003.

[9] S. Das, A. Chandrakasan, and R. Reif. Design tools for 3-D integrated circuits. In *Proc. ASP-DAC*, pages 53–56, 2003.

[10] S. Das, A. Chandrakasan, and R. Reif. Calibration of Rent's-rule models for three-dimensional integrated circuits. *IEEE TVLSI*, to appear.

[11] M. Wang, X. Yang, and M. Sarrafzadeh. Dragon2000: Fast standard-cell placement for large circuits. In *Proc. ICCAD*, pages 260–263, 2000.

[12] A. E. Caldwell, A. B. Kahng, and I. L. Markov. Can recursive bisection alone produce routable placements? In *Proc. 37th DAC*, pages 477–482, 2000.

[13] A. E. Dunlop and B. W. Kernighan. A procedure for placement of standard cell VLSI circuits. In *IEEE TCAD*, pages 92–98, 1985.

[14] C. Ababei, N. Selvakkumaran, K. Bazargan, and G. Karypis. Multi-objective circuit partitioning for cutsize and path-based delay minimization. In *Proc. ICCAD*, 2002.

[15] W. E. Donath *et al.* Timing driven placement using complete path delays. In *Proc. 27th DAC*, pages 84–89, 1990.

[16] C.-H. Tsai and S. Kang. Standard cell placement for even on-chip thermal distribution. In *Proc. ISPD*, pages 179–184, 1999.

[17] H. Thon. Three red-hot boxed cooler alternatives for the Athlon XP3200+. http://www.tomshardware.com/cpu/20030917/boxed_cooler-04.html, September 2003.

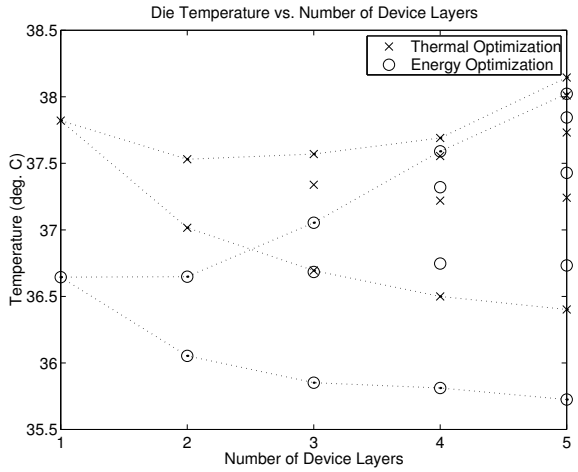

Figure 10: Die temperature of the FFT datapath vs. number of wafers (fixed-die case).

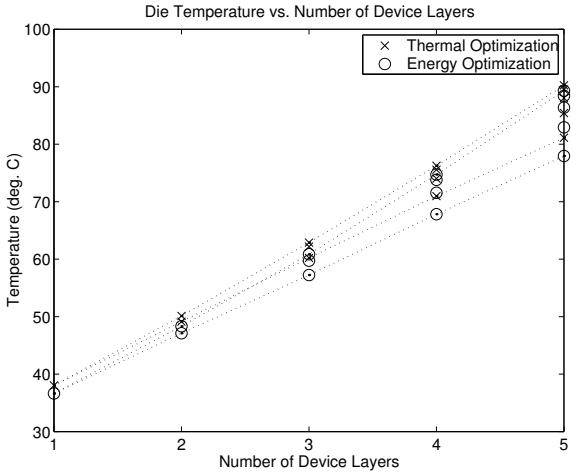

Figure 13: Die temperature of the FFT datapath vs. number of wafers (scaled-die case).

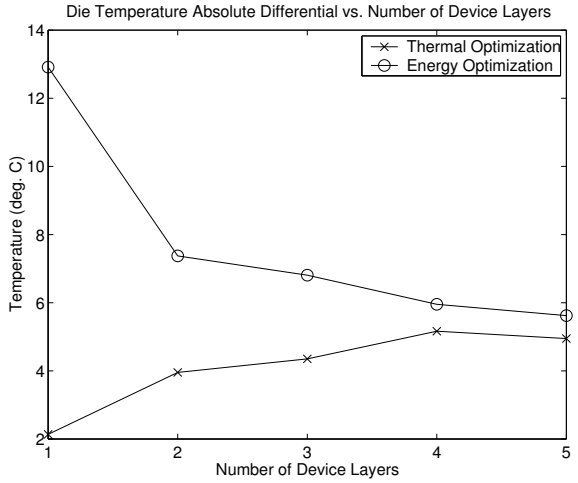

Figure 11: Absolute temperature differential of the FFT datapath vs. number of wafers (fixed-die case).

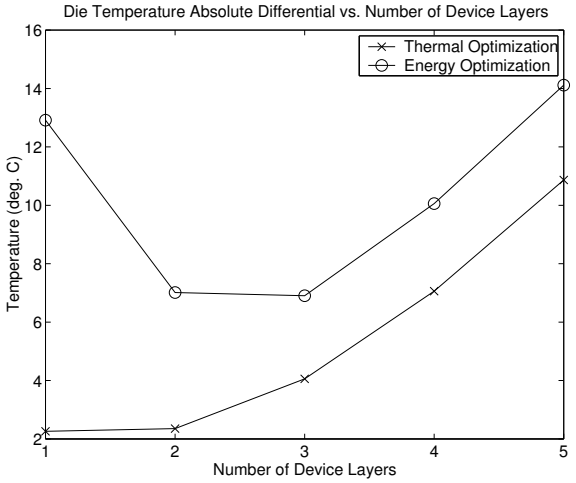

Figure 14: Absolute temperature differential of the FFT datapath vs. number of wafers (scaled-die case).

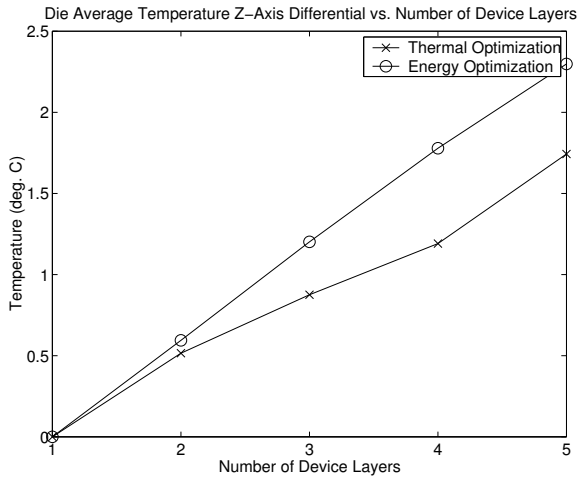

Figure 12: Average-temperature z-axis differential of the FFT datapath vs. number of wafers (fixed-die case).

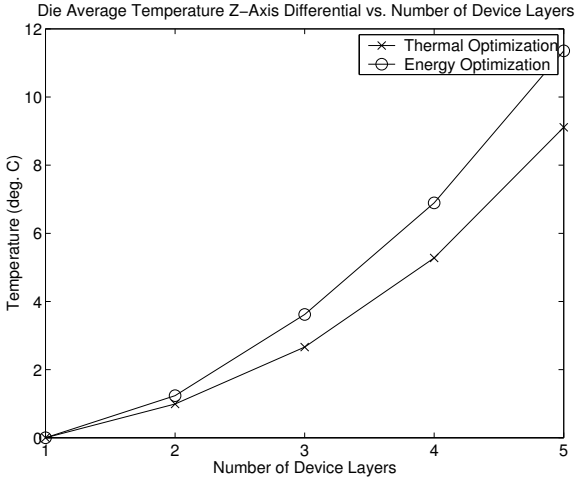

Figure 15: Average-temperature z-axis differential of the FFT datapath vs. number of wafers (scaled-die case).