Design Space Exploration for 3-D Cache

Yuh-Fang Tsai, *Member, IEEE*, Feng Wang, *Student Member, IEEE*, Yuan Xie, *Senior Member, IEEE*, Narayanan Vijaykrishnan, and Mary Jane Irwin, *Fellow, IEEE*

***Abstract*—As technology scales, interconnects have become a major performance bottleneck and a major source of power consumption for sub-micro integrated circuit (IC) chips. One promising option to mitigate the interconnect challenges is 3-D ICs, in which a stack of multiple device layers are put together on the same chip. In this paper, we explore the architectural design of cache memories using 3-D circuits. We present a delay and energy model 3-D cache delay-energy estimation tool (3D-Cacti) to explore different 3-D design options of partitioning a cache. The tool allows partitioning of a cache across different device layers at various levels of granularity. The tool has been validated by comparing its results with those obtained from circuit simulation of custom 3-D layouts. We also explore the effects of various cache partitioning parameters and 3-D technology parameters on delay and energy to demonstrate the utility of the tool.**

***Index Terms*—3-D integrated circuits (ICs), cache design, delay/energy simulator.**

I. INTRODUCTION

INTERCONNECTS have become a major performance bottleneck and a major source of power consumption for deep sub-micro integrated circuit (IC) chips. Consequently, interconnect centric design methods and technology improvements are critical to the chip industry. While there have been significant interconnect technology improvements over the last few years, such as the use of copper and low-K dielectric, the industry is striving for additional improvements. Various technologies have been actively explored to address the interconnect problem, such as the use of packet-based on-chip communication networks [3], the use of angular wires instead of Manhattan routing, and the design of 3-D chips [9].

A 3-D chip is a stack of multiple device layers with direct vertical interconnects tunneling through them. Fig. 1 shows a conceptual two-layer 3-D IC [11] (more details will be explained later). The direct vertical interconnects are called *inter-wafer vias*, or *through silicon vias (TSV)*. Compared to a traditional 2-D chip design, one of the most important benefits of a 3-D chip over a traditional 2-D design is the reduction on global interconnects. Joyner *et al.* [14] have shown that 3-D architectures reduce wiring length by a factor of the square root of the number of layers used. For example, the wiring length for a four-layer 3-D IC would be, on average, 50% of the wiring length for a 2-D IC chip, as illustrated in Fig. 2. The reduction of wire length due to 3-D integration can bring the following two obvious benefits.

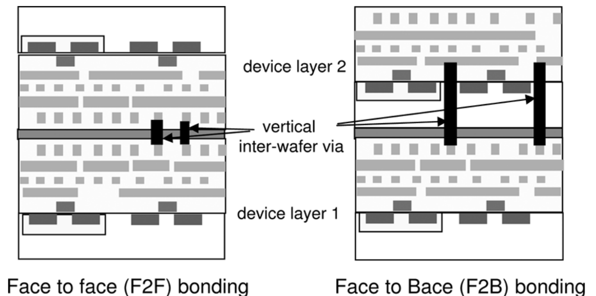


Fig. 1. Conceptual example of the implementation of 3-D IC using through via: two device layers are stacked together, using either F2F wafer bonding or F2B wafer bonding, with direct vertical interconnects tunnelling through them [11].

- ***Performance Improvement***. Higher performance can be achieved due to reduced average interconnect length and the critical path length. For example, Intel [20] has demonstrated that by targeting the heavily pipelined wires, the pipeline modifications resulted in approximately 15% improved performance, when the Intel Pentium-4 processor was folded onto two-layer 3-D implementation. Vaidyanathan *et al.* [38] also shown that the 3-D arithmetic unit design can achieve around 10%–30% delay reduction due to the wire length reduction.
- ***Power Reduction***. Interconnect power consumption becomes a large portion of the total power consumption as technology scales [24]. The reduction of the wire length translates into the power saving in 3-D IC design. For example, in the 3-D Intel Pentium-4 implementation [20], the wire reduction in the clock grid resulted in a 15% power reduction.

In addition to the performance and power benefits, other benefits of 3-D ICs include: 1) higher packing density and smaller footprint due to the addition of a third dimension to the conventional 2-D layout and 2) support for realization of mixed-technology chips.

Prior efforts have focused on developing different fabrication techniques involved in stacking multiple device layers and in forming the interconnects between layers. There are various 3-D technologies that have been explored in the past, including wire bonded, microbump, contactless (capacitive or inductive), and through via vertical interconnect. A comprehensive comparison of all these approaches have been described by Davis *et al.* [10]. Among these different approaches, through-via interconnect has the potential to offer the greatest vertical interconnect density and is the most promising one. In this paper, we consider two different styles of 3-D through via vertical interconnect

Manuscript received April 12, 2006; revised March 29, 2007. This work was supported in part by the National Science Foundation (NSF) under CAREER Award 0643902, NSF CCF 0702617, and NSF CRI 0202007.

The authors are with Computer Science Engineering Department, Pennsylvania State University, University Park, PA 16802 USA (e-mail: yuanxie@cse.psu.edu).

Digital Object Identifier 10.1109/TVLSI.2007.915429

1063-8210/$25.00 © 2008 IEEE

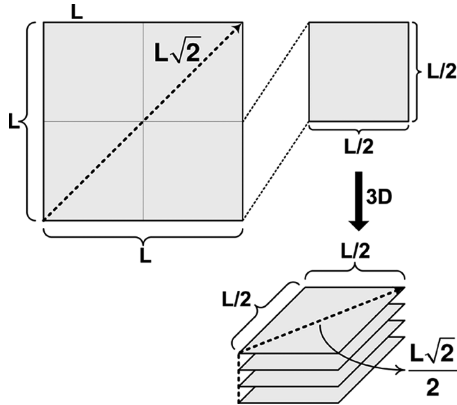


Fig. 2. Wire length reduction in 3-D IC: global interconnect scales in length as the square root of the number of stacked layers [14].

schemes: the top-down approach similar to wafer-bonding technology [9] and the bottom-up approach similar to multi-layer buried structures (MLBS) technology [15] described as follows.

- ***Bottom-Up Approach***. This approach involves sequential device process. An example of this approach is MLBS technology [15]. The front-end processing (to build the device layer) is repeated on a single wafer to build multiple active device layers before the back-end processing (to build the interconnects among devices).
- ***Top-Down Approach***. In this approach, each active device layers are processed separately, using conventional fabrication techniques. Then multiple device layers are assembled to build up 3-D IC. A typical example of this approach is *wafer bonding* [9]. After each wafer is processed individually, wafers can be bonded face-to-face (F2F) or face-to-back (F2B) (see Fig. 1). The through wafer via in F2F wafer-bonding does not go through thick buried silicon layer and can be fabricated with smaller via sizes. However, for 3-D ICs with more than two active layers, F2B wafer-bonding technology is necessary. Compared to the bottom-up approach (such as MLBS), the top-down approach (such as wafer-bonding) integration requires minimal change on the manufacture process steps, and therefore is more promising to become the main stream 3-D integration technology.

One of the key technology parameters in 3-D integration is the size of the vertical 3-D via (through silicon via or TSV), which provides connections between different active device layers and can have great influence on the following architecture partitioning strategies.

- In wafer-bonding, the dimensions of the 3-D via are not expected to scale at the same rate as feature size, because wafer-to-wafer alignment tolerances during bonding pose limitations on the scaling of the vias [4]. Current dimensions of 3-D via sizes vary from 1 μm by 1 μm to 10 μm by 10 μm [9], depending on the technology and manufacturing process used. Very recently, IBM has managed to reduce the pitch to a state-of-the-art 0.2 $\times$0.2 μm^2 using silicon-on-insulator (SOI) technology [40]. However, despite the decreasing via sizes, the via density does not scale aggressively due to wafer alignment and reliability concerns [4]. This via density limitation can hinder partitioning a design at very fine granularity across multiple device layers.
- The MLBS provides more flexibility in vertical 3-D connection because the vertical 3-D via can potentially scale down with feature size due to the use of local poly-Si wires for connection [9]. However, wafer-bonding requires fewer changes in the manufacturing process and is more popular in industry [18], [22] than MLBS technology.

Even though 3-D manufacture technology is becoming feasible, 3-D integration technology will not be commercially viable without the support of EDA tools and methodologies that allow architects and circuit designers to develop new architectures or circuits using this technology. To efficiently exploit the benefits of 3-D technologies, design tools, and methodologies to support 3-D designs are imperative. EDA design tools that are essential for 3-D IC design adoptions can be divided into two different categories; early design analysis tools and physical design tools. In this paper, we describe 3-D cache delay-energy estimation tool (3D-Cacti), which explores the architectural design space of cache memories using 3-D structures at the early design stage.

The regular structure and long wires in a cache make it one of the best candidates for 3-D designs. For example, Black *et al.* [45] have shown the benefit of 3-D cache design for Intel microprocessor, and Puttaswamy *et al.* [20] also demonstrated the 3-D cache design benefits in general. However, their 3-D cache designs are mainly custom design, and heavily depend on the designer's experience or may not be optimized for a particular design goal (such as performance or energy). To justify 3-D cost overhead, it is essential to study the benefits of 3-D integration at the early design cycle. The early analysis tools should facilitate the designers to study the tradeoffs among the number of layers, performance, energy, and power density.

In this paper, we investigate the architectural design of cache memories using 3-D integrated technology. Since interconnects dominate the delay of cache accesses, which determines the critical path of a microprocessor, the exploration of benefits from advanced technologies is particularly important. A tool[1] to predict the delay and energy of a cache is crucial as the timing profile and the optimized configurations of a cache design depend on the number of active device level available as well as the way the cache is partitioned into different active device layers. This paper examines possible partitioning approaches for caches designed using 3-D structures and presents a delay and energy model called 3D-Cacti to explore different options of partitioning a cache across different device layers.

The paper is organized as follows. Section II presents prior related work. Section III discusses possible 3-D partitioning approaches for a cache. Section IV presents a 3-D cache delay-energy estimator called 3D-Cacti, and the results of exploring the design space for 3-D cache are presented in Section V. Section VI concludes this paper.

II. Related Work

The prior work related to this paper can be divided into three main groups: system-level analysis of 3-D IC, 3-D physical de-

[1][Online]. Available: http://www.cse.psu.edu/~yuanxie/3d.html

sign tools, and architectural-level exploration. In the following paragraphs, we give a brief overview of these related works.

- ***System-Level Analysis of 3-D ICs***
  Early work on 3-D integrated circuits has been in the system-level modeling and analysis. To predict the reduction on wire-length and the improved performance in 3-D ICs, numerical models have been developed for various forms of 3-D integration technology [2], [14], [34]. Since interconnect power consumption is becoming an important portion of the total power consumption, the opportunities for reducing power dissipation using 3-D integration were also investigated [13]. The impact of the through via density on the performance was investigated [25]. The energy, thermal, and performance of 3-D designs under a given timing constraint were also studied [8]. Analysis of applying 3-D technology on FPGA design were conducted [17], [32]. These works have conducted a comprehensive system-level analysis of the benefits using 3-D integration, including the reduction of the wiring length, improved performance, and reduced power consumption. However, it also shows that the increased power density (due to stacking) can cause higher temperature [30], [33], which is often considered a major hindrance for the adoption of 3-D integration.
- ***Physical Design Tools for 3-D ICs***
  To efficiently exploit the benefits of 3-D technologies, design techniques and methodologies to support 3-D designs are imperative. 3-D IC design is fundamentally related to the topological arrangement of logic blocks. Therefore, physical design tools for 3-D circuits are important for the adoption of 3-D technology. New placement and routing tools are necessary to optimize 3-D circuits to take full advantages of the additional floorplanning/placement/routing dimension. As shown by system-level analysis [30], [33], a major concern in the adoption of 3-D technology is the increased on-chip temperature. Therefore, physical design tools for 3-D circuits have to be thermal aware. Recent efforts have focused on developing tools for supporting custom 3-D layouts and placement tools [9]. In [11], the technology and testing issues were surveyed and a physical design framework for 3-D ICs was presented. Thermal-driven floorplanning/placement/routing algorithms for 3-D ICs have been proposed by various groups [1], [7], [44]. To help mitigate thermal issues, thermal vias (the vertical inter-wafer vias which are only for heat conducting purpose and do not carry electrical signals) can be inserted during floorplanning and placement steps [39], [43]. Another design metric, reliability, was also considered during 3-D physical design steps [23], [35].
- ***Architectural Design Exploration and Evaluation***
  In mircoarchitecture level, a number of approaches have been proposed to explore the design space in 3-D microprocessors [5], [22], [27]. Black *et al.* [5] provided an overview of the potential benefits of designing an Intel IA32 processor in 3-D technology, even though the design details for each component were not disclosed. Loh *et al.* [27]–[29] investigated the 3-D implementation of important components in microprocessors, such as instruction schedulers, register files, and arithmetic units. Specific to 3-D memory design, a 3-D shared memory is fabricated. In this shared memory system design, six memory modules are distributed into three device layers and high data bandwidth is achieved by connecting broadcast bus in both horizontal and vertical directions (3-D vertical interconnects). As for cache design, Loh *et al.* [26] have shown a custom 3-D implementation of caches using F2F stacking. However, it did not fully explore all design options and thus the design itself may not be the optimal one.

Regular structure and long wires in a cache make it one of the best candidates for 3-D design. Also, as reported by Loh *et al.* [26], implementing only the caches in 3-D, without accounting for possible benefits from implementing other components of the processor in 3-D, can already result in a 12% instruction per cycle (IPC) gain for an Alpha-like microprocessor. To justify 3-D cost overhead, it is essential to study the benefits of 3-D cache design in the early design cycle. A tool to predict the delay and energy of a cache is crucial to explore the tradeoffs between different design goals (such as performance and energy) among many possible cache configurations. This paper examines possible partitioning approaches for caches designed using 3-D structures and presents a delay and energy model called 3D-Cacti [37] to explore different options of partitioning a cache across different device layers. A closely related work is another tool called PRACTICS developed by Zeng *et al.* [41]. However, their 3-D cache prediction tool did not consider leakage energy and thermal issues, which are important in 3-D cache design for technology nodes below 100 nm.

III. 3-D Cache Partitioning Strategies

In this section, we first give a brief illustration of the cache structure, and then describe different approaches to partition a cache into multiple device layers. A combination of different partitioning at various granularity discussed in this section is also possible.

A. Cache Structure

The structure of a cache contains a tag array and a data array. A portion of the address bits are used to index the corresponding set in the tag and data array. Fig. 3 shows the data array of a 32 kB cache (only data array is shown). Next, the tags and data of the different blocks belonging to a set are read. The tags read from all the blocks are compared against the tag portion of the incoming address. The indication of a match from the comparator output is used to enable the output driver of the corresponding block's data from the data array.

Neither the tag nor the data arrays are monolithic structures; the wordlines and bitlines of the memory array are divided into $Ndwl$ and $Ndbl$ parts resulting in $Ndwl \times Ndbl$ sub-arrays (blk0–blk7 in Fig. 3). This partitioning is effective in reducing the access times and power consumption. Since the dimensions of the tag and data arrays are different, they are typically partitioned differently. In a 3-D structure, we can extend this partitioning approach to divide bitlines and wordlines across different device layers. We will refer to this methodology as sub-array level partitioning in the following sections.

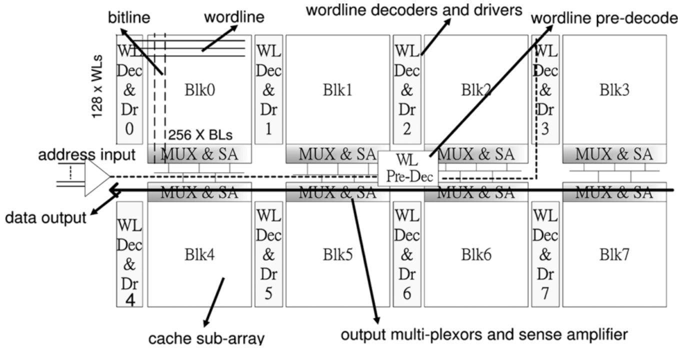


Fig. 3. Example layout of a 2D cache: each Blki is a sub-array with 128 wordlines and 256 bitlines. Note only data array is shown.

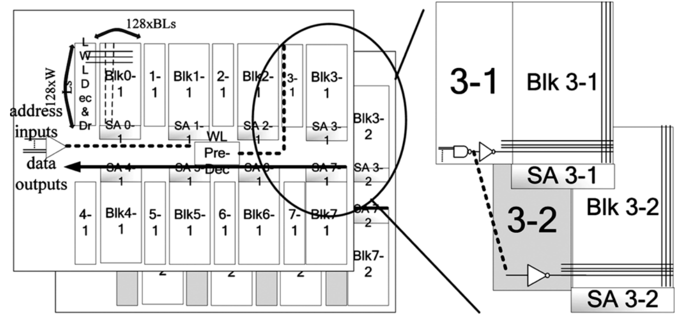


Fig. 4. 3-DWL approach: each sub-array is partitioned across wordline and mapped into two active device layers.

In addition to influencing the design of individual sub-arrays, the use of 3-D structures can also help to reduce the delays due to global interconnects in the cache. One of the global interconnects are the incoming address inputs to the cache that are sent to a predecoder, which is placed in the center of the sub-arrays. The predecoded address signals then traverse in an H-tree format to the local decoders of the sub-arrays. The local decoders in turn drive the corresponding word line drivers. Other global signals include the select signals for driving the output buffers of the data array, the wires from output driver to the edge of the array, and the select lines for write and multiplexer control. All these global signals should benefit from a smaller footprint through the use of 3-D technology. Global clock wiring will also benefit from 3-D cache design as it travels shorter distance, even though in our evaluation we do not account for the benefit of the clock network.

B. SRAM Cell Level Partitioning

The finest granularity of partitioning a cache is at the SRAM cell level. At this level of partitioning, any of the six transistors of an SRAM cell can be assigned to any layers. For example, the pull-up pMOS transistors can be in one device layer, while the access transistors and the pull-down nMOS transistors can be in another layer. The benefits of cell level partitioning include the reduction of footprints for the cache arrays and, consequently, the routing distance of global signals discussed. The number and complexity of the peripheral circuits remain the same as a conventional 2-D cache designs.

However, the feasibility of partitioning at this level is constrained by the 3-D via size and via density as compared to the SRAM cell size. Assuming a limitation that the size of 3-D vias cannot be scaled less than 1 μm by 1 μm, a 3-D via has a comparable size to that of a 2-D 6T SRAM cell in 180-nm technology and is much larger than a single cell in 70-nm technology. Note that for the 70-nm technology, the size difference remains even when using a "wide cell" topology for the SRAM cell in deep submicrometer design to alleviate process difficulties of small feature sizes [42]. Consequently, when the 3-D via size does not scale with feature size as currently in wafer-bonding, partitioning at the cell level is difficult in future technology nodes. In contrast, partitioning at SRAM cell level will continue to be feasible in technologies such as MLBS, because no limitations are imposed on via scaling with feature size. Availability of such technologies makes it possible to partition the cache at the granularity of individual cache cells [16]. However, it should be noted that even if the size of a 3-D via can be scaled to as small as a nominal contact in a given technology, the total SRAM cell area reduction (as compared to a 2-D cache design) due to the use of additional layers is limited, because metal routing and contacts occupy a significant portion of the 2-D SRAM cell area [42]. Consequently, partitioning at higher levels need to be explored. Furthermore, wafer-bonding requires fewer changes in the manufacturing process and is more popular in industry [18], [22] than MLBS technology. Therefore, our 3-D cache design space exploration is mainly focused on coarse level partitioning using wafer-bonding technology.

C. Sub-Array Level Partitioning

At this level of partitioning, individual sub-arrays in the 2-D cache are partitioned across multiple device layers. The partitioning at this granularity reduces the footprint of cache array and routing length of global signals. However, it also changes the complexity of the peripheral circuits. In our research, we consider two options of partitioning sub-arrays into multiple layers: 3-D divided wordline (3-DWL) approach and 3-D divided bit line approach (3-DBL).

3-DWL: In this partitioning strategy, wordlines in a sub-array are divided and mapped onto different active device layers (see Fig. 4). The corresponding local wordline decoder of the original wordline in 2-D sub-array is placed on one layer and is used to feed the wordline drivers on different layers through the 3-D vias. Instead of a single wordline driver as in the 2-D case, we have multiple word line drivers in the new design for each layer. The duplication overhead is offset by the resized drivers for a smaller capacitive load on the partitioned word line. Further, the delay time of pulling a wordline decreases as the number of pass transistors connected to a wordline driver is smaller. The delay calculation of the 3-DWL also accounts for the 3-D via area utilization. The area overhead due to 3-D vias is small compared to the number of cells on a wordline.

Another benefit from 3-DWL is that the distance of the address line from periphery of the core to the wordline decoder decreases as the number of device layers increases. Similarly, the routing distance from the output of predecoder to the local decoder is reduced. The length of select lines for the writes and MUXes, as well as the wires from the output drivers to the periphery is also reduced.

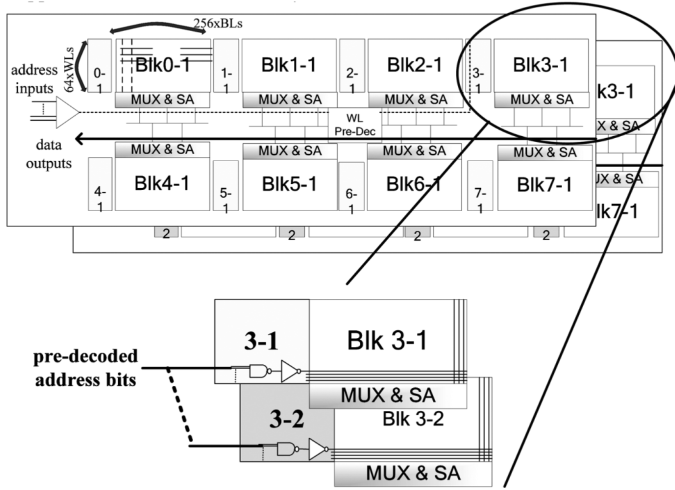


Fig. 5. 3-DBL approach: each sub-array is partitioned across wordline and mapped into two active device layers.

TABLE I
DESIGN PARAMETERS FOR 3D-CACTI

	Definition	Effect
Ndbl	number of cuts on a cache to divide bitlines	1. bitline length in each sub-array 2. number of sense amplifier 3. size of wordline driver 4.decoder complexity 5. multiplexors complexity in data output path
Ndwl	number of cuts on a cache to divide wordlines	1. wordline length in each sub-array 2. number of wordline driver 3. decoder complexity
Nspd	number of sets connected to a wordline	1. wordline length in each sub-array 2. size of wordline driver 3. multiplexors complexity in data output path
Nx	number of 3D partitions by dividing wordlines	1. wordline length in each sub-array 2. size of wordline driver
Ny	number of 3D partitions by dividing bitlines	1. bitline length in each sub-array 2. complexity in multiplexors in data output path

3-DBL: This approach is akin to the 3-DWL approach and applies partitioning to the bitlines of a sub-array (see Fig. 5). The bitline length in the sub-array as well as the number of pass transistors connected to a single bitline is reduced, which can facilitate faster switch of the bitline. In the 3-DBL approach, the sense amplifiers can either be duplicated across different device layers or shared between the partitioned sub-arrays in the different layers. The former approach is more suitable for reducing access times while the latter is preferred for reducing number of transistors and leakage. In the latter approach, the sharing increases complexity of multiplexing of bitlines and reduces performance as compared to the former. Note that sharing the sense amplifier among multiple layers will increase the number of 3-D vias, because each bitline has to use one 3-D via to share the sense amplifier. This can potentially cause the via congestion problem, and may not be feasible for some 3-D integrations with much larger 3-D via (such as 10 μm by 10 μm), even though it is possible for much smaller 3-D via (such as IBM's 0.2 μm by 0.2 μm technology [4] or MLBS 3-D integration). On the other hand, when sense amplifiers are duplicated for each die in the stack, the via congestion problem can be avoided, at the expense of more transistors and extra leakages. Similar to 3-DWL, the length of the global lines are reduced in this scheme.

IV. 3D-CACTI

In order to explore the 3-D cache design space, we develop 3D-Cacti. Our tool is built on top of the Cacti 3.0 2-D cache tool (cache access and cycle time) [36]. In addition to the 3-D enhancements, we improve the models used for technology scaling and leakage power in the base 2-D case. 3D-Cacti searches for the optimized configuration that provides the best delay, power, and area efficiency tradeoff according to the cost function for a given number of different 3-D partitions. The cost function for each cache configuration i is defined as [36]

$$\mathrm{cost}_i = \frac{\mathrm{energy}_i}{\mathrm{energy}_{\max}} * W_e + \frac{\mathrm{accesstime}_i}{\mathrm{accesstime}_{\max}} * W_t + \frac{1}{\mathrm{areaefficency}_i} * W_{ae} + \frac{\mathrm{aspectratio}}{\mathrm{aspectratio}_{\max}} * W_{ar} \quad (1)$$

where W_e, W_t, W_{ae}, and W_{ar} are the weight of energy, delay, area efficiency, and array aspect ratio, respectively.

A. Implementation of 3D-Cacti

In this section, we only emphasize the portions of 3D-Cacti that are different from the original Cacti 3.0. We have grouped these changes into delay models, layout parameters, and technology parameters.

Delay Models: To model the resistance and capacitance of the 3-D vias, we add the resistance–capacitance *(RC)* delay to implement the intra-array partitioning. The resistance of 3-D via is estimated to be $10^{-8}\ \Omega-\mathrm{cm}^2$ based on actual resistance measurement [6], and the capacitance is estimated as the capacitance of a 1 μm by 1 μm contact using top metal layer and the height of the interlayer via is assumed to be 10 μm. With technology scaling, repeater insertion for long wires is necessary. We also enhance the wire delay model by implementing repeater insertion for long interconnects [31].

Layout Parameters: Several configuration parameters are used in the original Cacti to divide a cache into sub-arrays to achieve delay, energy, and area efficiency tradeoffs. In our implementation, two additional parameters, Nx and Ny, are added to model the sub-array level 3-D partitions.

The additional effects of varying each parameter other than the impact on length of global routing signals are listed in Table I. Note that the tag array is optimized independently of the data array and the configuration parameters for tag array ($Ntwl$, $Ntbl$, and $Ntspd$) are not shown in this table. Fig. 6 shows an example of how these configuration parameters are used in 3D-Cacti affect the cache structure. The cell level partitioning approach is implicitly simulated using a different cell width and height within Cacti.

Technology Parameters: The scaling of the delay in transistors is different from that in wires. The original Cacti is built based on the technology parameters of 0.8 μm technology. To estimate the delay and energy for smaller technologies, instead of applying the linear scaling on the final delay and energy numbers as in Cacti, we apply more accurate scaling rules derived

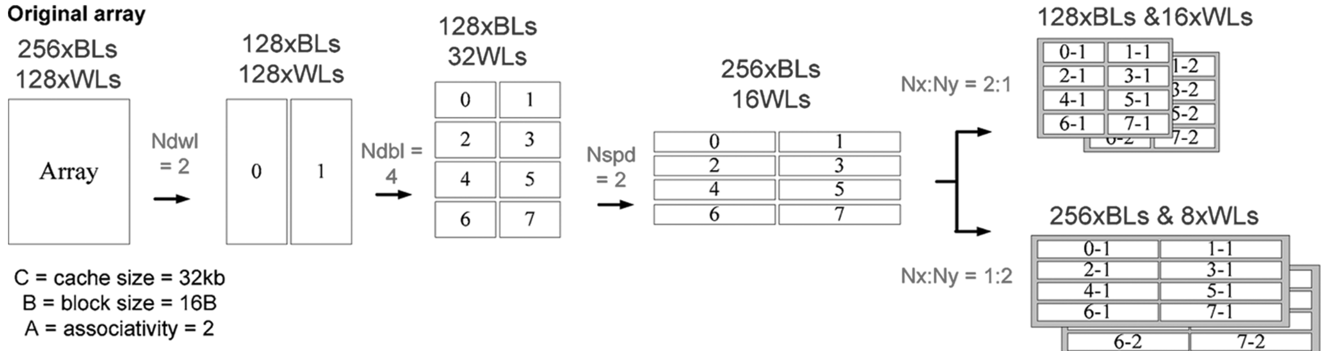


Fig. 6. Example showing how each configuration parameter affects a cache structure. Each box is a sub-array associated with an independent decoder.

from [12] on each individual technology parameter. We also assume the use of copper interconnect for technologies smaller than 0.18 μm and account for the fact that aspect ratio of 6T SRAM is getting smaller as technology scales. We also adopt the "wide-bit" cell design for technologies smaller than 70 nm according to the SRAM design fabricated in 65 nm [42]. We augment leakage models in the cache to account for leakage energy [21]. Different from [21], we use the transistor sizes scaled from original Cacti and insert repeaters where there are long wires or large loadings. The sizes of all transistors and repeaters are coupled with the configuration information to account for the transistor counts to estimate total leakage power.

B. Temperature Estimation

One of the major concerns for the adoption of 3-D technology is the increased power density due to stacking. Higher power density causes temperature increases, which can affect the performance and leakage power. In order to facilitate accurate estimation of performance and energy, a compact thermal model is needed to provide the temperature profile. Numerical computing methods [such as finite difference method (FDM)] are very accurate but computationally intensive. Skadron *et al.* proposed a thermal model called Hotspot,[2] which is based on lumped thermal resistances and thermal capacitances. It is more efficient than the prior low-level approaches since the variances at temperature are tracked at a granularity of functional block level. We perform temperature estimates using a 3-D IC thermal analysis tool called HS3-D [19]. The thermal analysis tool is first fed with the power and area information assuming operating at room temperature and the estimated temperature is then feedback to 3D-Cacti to account for the thermal difference due to stacking. This temperature-leakage inter-dependence is accounted for in 3D-Cacti. The increasing temperature due to the stacking [8] is factored into the leakage estimates.

C. Validation of 3D-Cacti

In order to validate 3D-Cacti, we implement the layouts of a 32 kB two-way set associative cache with 16 byte blocks in TSMC 0.18-μm technology with a publicly available design and layout extraction tool called 3-DMagic [8], [9], for three different cases. These three cases include a 2-D layout, two-layer 3-D layouts employing 3-DWL (i.e., $Nx \times Ny = 2 \times 1$), and two-layer 3-D layouts employing 3-DBL (i.e., $Nx \times Ny = 1 \times 2$). For the 3-D design, the layout for each layer is designed separately and the 3-D vias is modeled as special contacts. The *RC* characteristics of 3-D vias were modeled based on parameters described in Section IV-A.

TABLE II
REDUCTIONS IN DELAY AND ENERGY ESTIMATION FROM 3D-CACTI AND HSPICE SIMULATION RESULTS OF TWO-LAYER LAYOUT AS COMPARED TO A 2-D CACHE

Method	3DCacti 3DWL	HSPICE 3DWL	3DCacti 3DBL	HSPICE 3DBL
Performance Savings	17.75%	17.33%	6.15%	6.91%
Power Savings	14.95%	9.23%	15.71%	10.07%

The 3D-Cacti estimates of delay and power are compared against HSPICE simulation results from the layouts. The savings in delay and power of 3-DWL and 3-DBL as compared to the 2-D design estimated by 3D-Cacti and HSPICE are shown in Table II. We observe that the relative delay and power trends for these designs predicted by our model and the actual designs are similar. We validate the relative trends instead of absolute numbers as the underlying technology parameters used by TSMC 0.18-μm process by HSPICE and the scaled technology numbers used in 3D-Cacti for 0.18 μm are different. The normalized delay of each individual component of the cache for 3-DWL and 3-DBL is shown in Figs. 7 and 8, respectively. This demonstrates that each individual component delay modeled by 3D-Cacti also matches with HSPICE simulation results.

V. DESIGN EXPLORATION USING 3D-CACTI

In this section, we explore various 3-D partitioning options of caches using 3D-Cacti to understand the impact on delay and power. Furthermore, we investigate the influence of system requirements, numbers of active device layers, and technology scaling on the 3-D cache performance. Note that the data presented in this section is for 70-nm technology, unless otherwise stated.

[2][Online]. Available: http://www.lava.cs.virginia.edu/HotSpot/

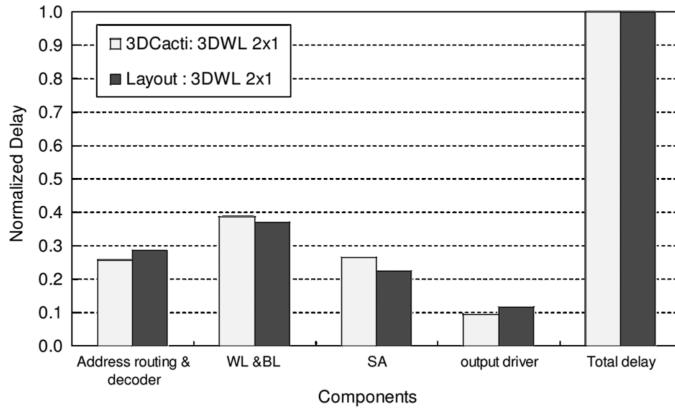


Fig. 7. Component-wise comparisons of the delay estimation from 3D-Cacti and HSPICE simulation results of two-layer layout for 3-DWL. Cache size is 32 kB.

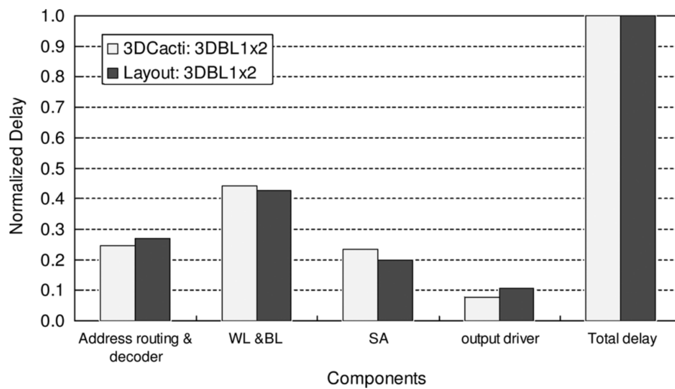


Fig. 8. Component-wise comparisons of the delay estimation from 3D-Cacti and HSPICE simulation results of two-layer layout for 3-DBL. Cache size is 32 kB.

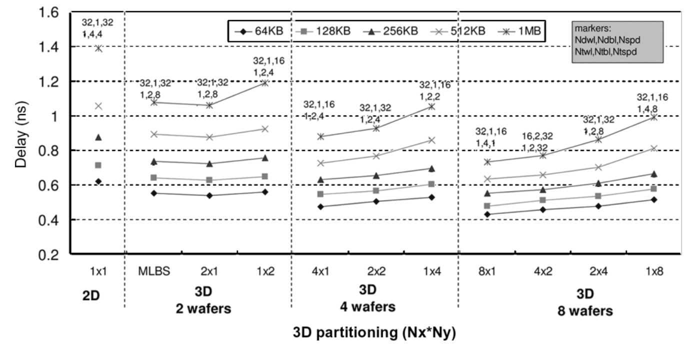


Fig. 9. Access time for different partitioning when setting the weight of delay W_t higher. Data of caches of associativity $= 4$ are shown.

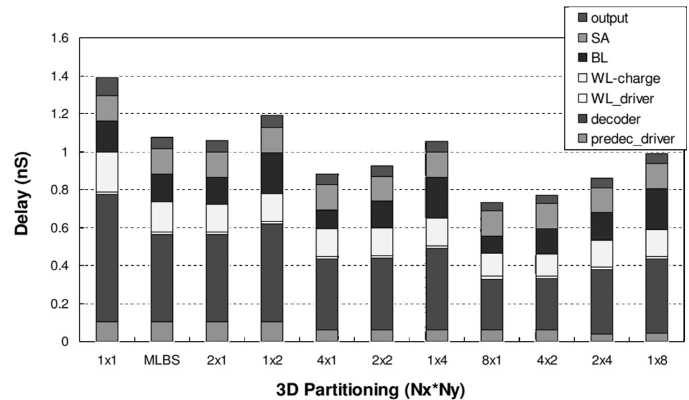


Fig. 10. Access time breakdown of a 1 MB cache corresponding to the results shown in Fig. 9.

A. Impact of 3-D Partitioning on Cache Performance and Energy

First, we set the weights in the cost function (1) to be $(W_e : W_t : W_{ae} : W_{ar}) = (1 : 1000 : 1 : 1)$, so that the performance is the major design goal. We explore the design space to find the best configurations for various degrees of 3-DWL and 3-DBL in terms of delay. Figs. 9 and 12 show the access delay and energy consumption per access for four-way set associative caches of various sizes and different 3-D partitioning settings. Recollect that $Nx(Ny)$ in the configuration refers to the degree of 3-DWL (3-DBL) partitioning. First, we observe that the delay is reduced as the number of layers increases. To have a better understanding of the reason for the delay and energy improvements, we study the access time and energy breakdown for each individual component, which are shown in Figs. 10 and 13, we observe that the reduction in global wiring length of the decoder is the main reason of benefits from 3-D design. We also observe that for the two-layer case, the partitioning of a single cell using MLBS provides delay reduction benefits similar to the best sub-array level partitioning technique as compared to the 2-D design. Note that in Fig. 13, the energy consumption by sense amplifiers dominates, due to the power-inefficient design modeled by the original Cacti tool. If a power-conscious sense amplifier design is used, the 3-D benefit can potentially be much better, since the interconnect power would be a larger fraction of the total power consumption.

Another general trend observed for all cache sizes indicates that partitioning more aggressively using 3-DWL (i.e., having a larger Nx value) provides caches with smaller delay. For example, in the four-layer case, the configuration 4×1 (which means $N_x = 4$ and $N_y = 1$) has a delay that is 16.3% faster than that of the 1×4 configuration for a 1-MB cache. We observed that the benefits from more aggressive 3-DWL stem from the fact that, in the original 1-MB 2-D cache design, the global wires in the X direction has much longer length compared to the global wires in the Y direction [see Fig. 11(a)]. The reason that the optimal 1-MB 2-D cache has a "wide" shape is that shorter bitlines are favorable for delay minimization in the original Cacti tool, resulting wider sub-arrays with difference in wire lengths along X and Y directions. Consequently, when we apply 3-DWL to partition each sub-array by cutting wordlines, we can achieve significant reduction in critical global wiring lengths. Note that because 3D-Cacti is exploring partitioning across the dimensions simultaneously, some configurations can result in 2-D configurations that have wirelength longer in the Y directions [See Fig. 11(c)] as in the 1-MB cache 1×2 configuration for two layers. The 3-DBL helps in reducing the global wire length delays by reducing the Y direction length. However, it is still not as effective as the corresponding 2×1 configuration as both the bitline delays in the sub-array and the routing delays are longer (see Fig. 10). These trends are difficult to analyze without the help of a tool to partition across

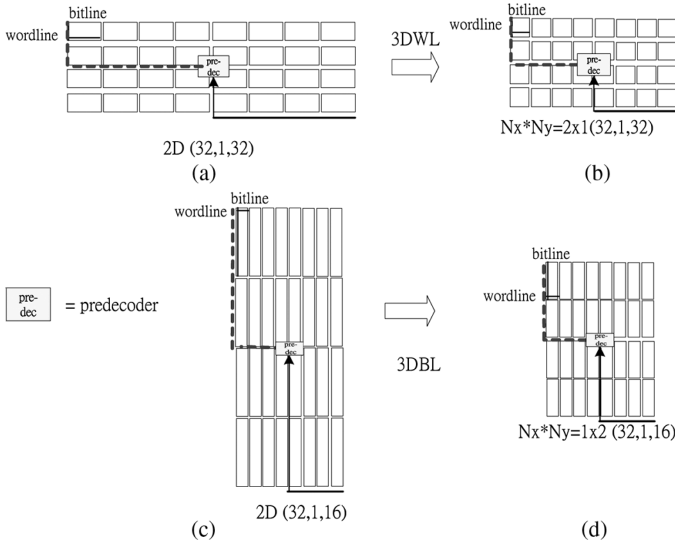


Fig. 11. Critical paths in 3-DWL and 3-DBL for a 1-MB cache. Dashed lines represent the routing of address bits from predecoder to local decoder while the solid arrow lines are the routing paths from the address inputs to predecoders.

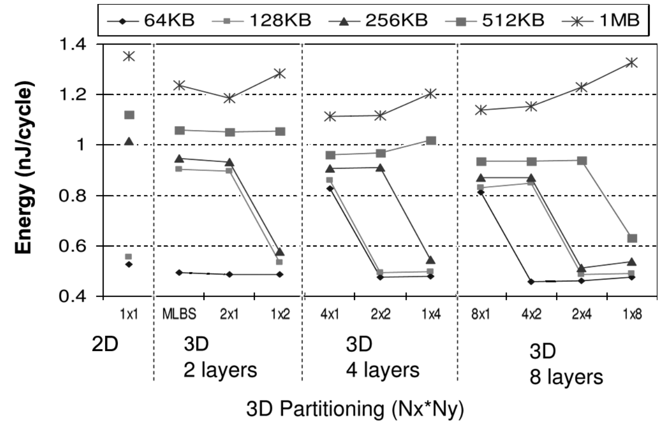


Fig. 12. Corresponding energy for different partitioning as shown in Fig. 9. Data of caches of associativity = 4 are shown.

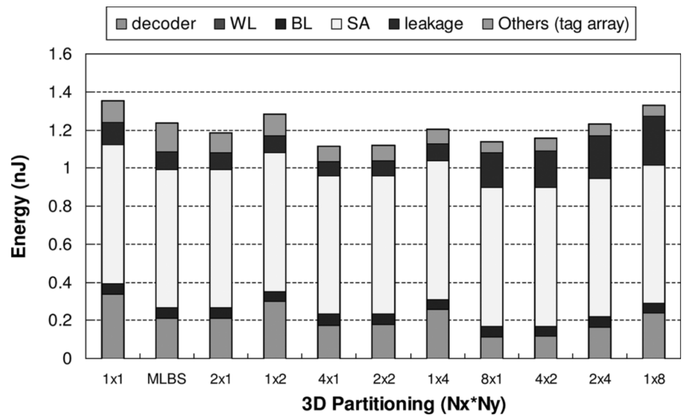


Fig. 13. Energy breakdown of a 1-MB cache corresponding to the results shown in Fig. 12.

multiple dimensions simultaneously. The energy estimation for the corresponding best delay configurations tracks the delay behavior in many cases. For example, the 1-MB cache energy behavior increases when moving from a 8 × 1 configuration to a 1 × 8 configuration. In these cases, the capacitive loading, which affects delays, also affects the energy trends. However, in some cases, the energy reduces significantly when changing configurations and does not track performance behavior. For example, for the 512-kB cache using eight-layers, the energy reduces when moving from 2 × 4 to 1 × 8 configuration. This stems from the difference in the number of sense amplifiers activated in these configurations due to the different number of bitlines in the each subarray for the different configurations, and the presence of the column decoders after the sense amplifiers. Specifically, the optimal $(Ndwl, Ndbl, Nspd)$ for the 512-kB case is (32,1,16) for the case in which $Nx \times Ny = 2 \times 4$, and (32,1,8) for the case in which $Nx \times Ny = 1 \times 8$, respectively. Consequently, the number of sense amplifiers activated per access for the latter case is half as much as that of the former case, resulting in a smaller energy.

Note that in Fig. 9, it is clear that for larger cache size, the delay improvement due to 3-D stacking is larger. However, for the corresponding energy consumption in Fig. 12, the trend is not clear. The major reason is that, in this experiment, we set the weights in the cost function (1) to be $(W_e : W_t : W_{ae} : W_{ar}) = (1 : 1000 : 1 : 1)$, such that the performance is the major design goal. Therefore, the fastest configuration may not be the most energy efficient one (for example, in 3-DBL approach, the tool may duplicate sense amplifiers to help access time, at the expense of larger leakage energy).

B. Varying Cost Function

The delay/energy savings achieved by 3-D partitioning depends on the tradeoff between delay, energy, and area efficiency. This tradeoff varies according to system requirements. To explore the impact of system requirements on the delay/energy savings, we conduct experiment on a 1-MB cache $(\text{associativity} = 4)$ for different system requirements. We set the weights in the cost function to be $(W_e : W_t : W_{ae} : W_{ar}) = (1 : 1000 : 1 : 1)$ and $(W_e : W_t : W_{ae} : W_{ar}) = (1000 : 1 : 1 : 1)$ for high performance systems and low power systems, respectively. The results are shown in Figs. 14 and 15. We can see that 3-DWL is more efficient than 3-DBL in terms of both delay and energy in high performance systems while 3-DBL is more effective in low power systems. This is because when optimizing for low power, the optimal configuration partitions a cache in a fashion that the routing in Y direction is longer than that in X direction, which is different from that in high performance systems. The optimal configuration with each layer changes across different 3-D partitioning approaches and system requirements. Further, the energy for the fast delay time configurations for different 3-D partitioning are quite different. Consequently, simultaneous exploration of different parameters along with desired weights for different cost functions is important in deciding the 3-D partitioning.

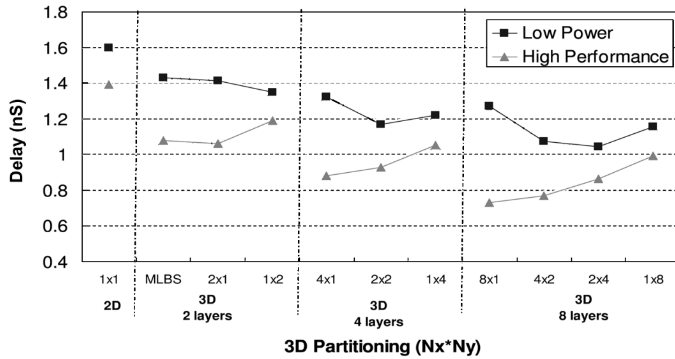


Fig. 14. Comparison of access time for different partitioning for high performance and low power systems. Data of associativity $= 4$ are shown.

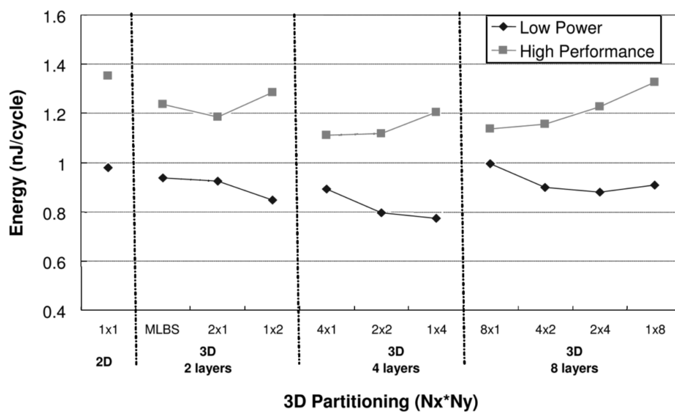


Fig. 15. Comparison of energy for different partitioning for high performance and low power systems. Data of associativity $= 4$ are shown.

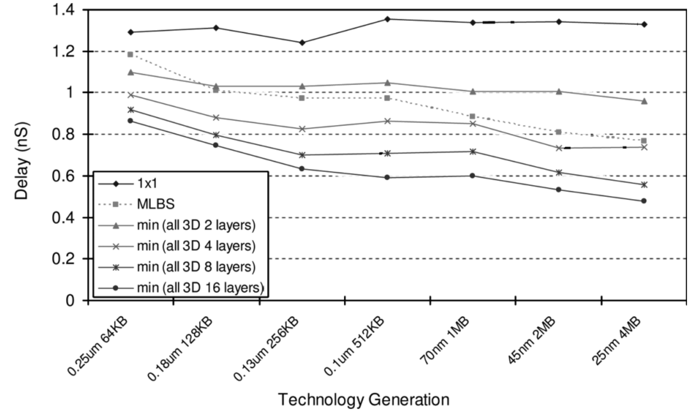


Fig. 16. Delay time across technology generations. Data shown are for set associative $(A = 4)$ cache.

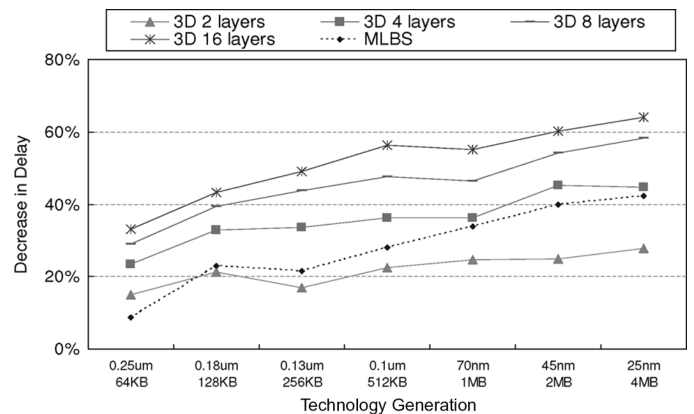


Fig. 17. Decrease in delay time comparing to 2-D cache (1×1 case) across technology generations. Data shown are for set associative $(A = 4)$ cache.

C. Impacts of Number of Active Device Layers and Technology Scaling

Figs. 16 and 17 show the access delay time and reduction in access delay of set associative cache $(A = 4)$ for various numbers of active device layers across technology generations, respectively. Note that we assume that the cache sizes double every technology generation to reflect the scaling trend, and the cache sizes for each technology node are shown in Figs. 16 and 17. The delay shown is the minimum delay achievable with any 3-D partitioning for a given number of active device layers. The delay reduction achieved in 70-nm technology is between 24.56% and 55.05% for using two active device layers to 16 layers. We observe that a significant delay reduction is obtained when moving from one layer to two device layers. The incremental reduction in delay with increasing number of layers becomes smaller as the gate delays become comparable to the interconnect delays when stacking more layers.

Stacking multiple layers of memory cells results in the increased power density, which can cause higher temperature [30], [33]. Thermal issue is often considered a major hindrance for the adoption of 3-D integration. Even though there are no more cache-only die, and the thermal behavior for the cache in a microprocessor is greatly affected by the higher power density microprocessor cores sitting next to the cache, we investigate the thermal trends for stacking caches, to provide some design guideline for microarchitecture designers who integrate cache design together with the rest of the system.

We adopt the package assumption of HS3-D tools [19]. A 1.8 cm by 1.8 cm by 500 μm silicon substrate sits adjacent to a 10-μm-thick silicon active layer. The substrate is connected to a 1-mm-thick 3 cm by 3 cm copper heat spreader by 75 μm of thermal interface material. The heat spreader is connected to a 6 cm by 6 cm aluminum heat sink with a 6.9-mm-thick base and 256 evenly distributed 2 mm $\times$ 2 mm pins, with an ambient temperature of 35 °C. With this package assumption, our results of a 1-MB cache in 45 nm show that chip temperature rises from 45 °C in 2-D chip to 50 °C, 65 °C, 92 °C, and 148 °C when stacking 2, 4, 8, and 16 layers of devices layers, respectively. Therefore, the increase in leakage power for stacking 8 and 16 active device layers are three times and six times, respectively, because SRAM cells become more leaky as temperature increases. Our results show that using two or four active device layers, which are more practical for wafer stacking, the leakage energy caused by thermal issue is not significant (5% and 18%, respectively). However, using more than eight device layers may not provide any reduction in total energy as temperature-dependent leakage is significant. Note that this analysis is based on the assumption of stacking-cache-only scenario. For a real system design, where a much hotter core sitting right next

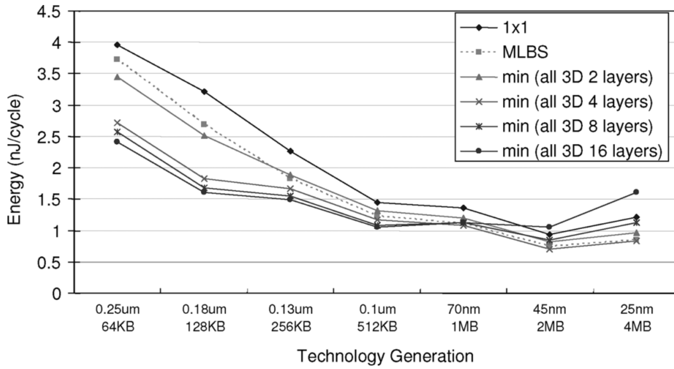


Fig. 18. Energy across technology generations. Data shown are of set associative $(A = 4)$ cache.

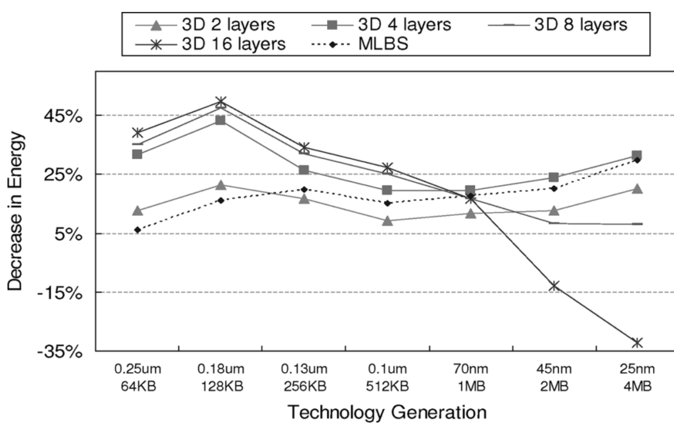


Fig. 19. Decrease in energy comparing to 2-D cache (1×1 case) across technology generations. Data shown are of set associative $(A = 4)$ cache.

to the cache, the thermal problem due to 3-D stacking is definitely more pronounced.

Figs. 18 and 19 show the minimum energy achieved for each 3-D partitioning and reduction in energy of set associative cache $(A = 4)$ for various numbers of active device layers across technology generations, respectively. This relationship between the reduction in energy and the number of active device layers is different from that observed for delay variation with number of layers. First, the reduction in energy is limited due to the fact that the power in sense amplifiers dominates overall power consumption and do not benefit from sub-array level 3-D partitioning. The energy savings in 25-nm technology are 20.1%, 31.38%, 7.91%, and −32.23% for 2, 4, 8, and 16 active device layers, respectively. The small energy savings for stacking eight device layers and increasing energy for stacking 16 device layers are caused by the increasing temperature due to stacking and thus increasing leakage energy.

The technology scaling trend of delay/energy of 3-D caches across technology generations can also be observed from Figs. 16–19. The delay time decreases with technology scaling. The percentage of reduction in delay increases across technology generations due to the increasing global wire delay. The savings in energy through 3-D partitioning is limited when leakage power is significant and higher temperature due to stacking is taken into account. The results in 45- and 25-nm technologies show that using more than eight device layers will not be beneficial in terms of energy due to the increased temperature and thus elevated leakage power.

Note that even though in our 3D-Cacti tool, we have delay as well as energy estimation, as technology scales, leakage starts to dominate, and leakage is strongly affected by chip temperature, which cannot be analyzed by only studying the standalone cache stacking. We believe that the most valuable part for 3D-Cacti is the cache access estimation during early design space exploration, while the energy and temperature estimation could provide a reference for the designers for later evaluation for the whole microprocessor design.

VI. Conclusion

3-D ICs are an attractive option to overcome the barriers in interconnect scaling, offering an opportunity to continue the CMOS performance trend. A tool to predict the delay and energy of a cache at the early design stage is crucial as the timing profile and the optimized configurations of cache depend on the number of active device level available as well as the way a cache is partitioned into different active device layers. In this paper, we explore the architectural design of cache memories using 3-D technologies. A cache delay and energy estimator called 3D-Cacti is proposed to explore different options to partition a cache across different active device layers. The estimator has been validated using actual designs. We observe that the savings in delay/energy through 3-D partitioning depends on the cache size, system requirements, the number of device layers, and technology nodes. Of course, the tool itself has some limitations. For example, the sense amplifier model is based on the old model from original CACTI, which is known to be a very power-inefficient design, and therefore has a large percentage of the energy in Fig. 13. If a power-conscious sense amplifier design is used, the 3-D benefit can potentially be much better, since the interconnect power would be a larger fraction of the total energy consumption. On the other hand, current model does not incorporate a lot of wire engineering and therefore for a cache design with enough wire engineering, the 3-D benefit may not be as large as the model predicts. After all, as an architectural level early analysis tool, 3D-Cacti can be used to predict relative benefits during architecture design space exploration.

Acknowledgment

The authors would like to thank IBM 3-D Program Managers K. Bernstein and A. Young for their invaluable help with the understanding of 3-D fabrication process.

References

[1] K. Balakrishnan, V. Nanda, S. Easwar, and S. K. Lim, "Wire congestion and thermal aware 3d global placement," in *Proc. Asia South Pacific Des. Autom. Conf. (ASPDAC)*, 2005, pp. 1131–1134.

[2] K. Banerjee, S. Souri, P. Kapur, and K. Saraswat, "3-D ICs: A novel chip design for improving deep-submicrometer inteconnect performance and system-on-chip integration," *Proc. IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.

[3] L. Benini and G. D. Micheli, "Networks on chips: A new SOC paradigm," *Computer*, no. 1, pp. 70–78, 2002.

[4] K. Bernstein, "Introduction to 3D integration," presented at the Tutorials Int. Solid State Circuits Conf. (ISSCC), San Francisco, CA, 2006.

[5] B. Black, D. W. Nelson, C. Webb, and N. Samra, "3D processing technology and its impact on IA32 microprocessors," in *Proc. IEEE Int. Conf. Comput. Des. (ICCD)*, 2004, pp. 316–318.

[6] K. N. Chen, A. Fan, C. S. Tan, and R. Reif, "Contact resistance measurement of bonded copper interconnects for three-dimensional integration technology," *IEEE Electron Devices Lett.*, vol. 25, no. 1, pp. 10–12, Jan. 2004.

[7] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," in *Proc. Int. Conf. Comput.-Aided Des. (ICCAD)*, 2004, pp. 306–313.

[8] S. Das, A. Chandrakasan, and R. Reif, "Timing, energy, and thermal performance of three-dimensional integrated circuits," in *Proc. 13th ACM Great Lakes Symp. VLSI (ISVLSI)*, 2004, pp. 338–343.

[9] S. Das, A. Fan, K. N. Chen, C. S. Tan, N. Checka, and R. Reif, "Technology performance and computer-aided design of three-dimensional integrated circuits," in *Proc. Int. Symp. Phys. Des. (ISPD)*, 2004, pp. 108–115.

[10] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon, "Demystifying 3D ICs: The pros and cons of going vertical," *IEEE Des. Test Comput.*, vol. 22, no. 6, pp. 498–510, Nov. 2005.

[11] Y. Deng and W. Maly, "2.5-dimensional VLSI system integration," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 6, pp. 668–677, Jun. 2005.

[12] G. Farland, "CMOS technology scaling and its impact on cache delay," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, 1997.

[13] J. Joyner and J. Meindl, "Opportunities for reduced power dissipation using 3D integration," in *Proc. Int. Test. Conf. (ITC)*, 2002, pp. 148–150.

[14] J. W. Joyner, P. Zarkesh-Ha, and J. D. Meindl, "A stochastic global net-length distribution for a three-dimensional system-on-a-chip (3D-SoC)," in *Proc. 14th Ann. IEEE Int. ASIC/SOC Conf.*, Sep. 2001, pp. 147–151.

[15] S. M. Jung *et al.*, "The revolutionary and truly 3-dimentional 25F2 SRAM technology with the smallest S3 cell, 0.16 μ^2 and SSTFF for ultra high density SRAM," in *VLSI Technology Dig. Techn. Papers*, 2004, pp. 228–229.

[16] Y. H. Kang, S. M. Jung, J. H. Jang, J. H. Moon, W. S. Cho, C. D. Yeo, K. H. Kwak, B. H. Choi, B. J. Hwang, W. R. Jung, S. J. Kim, J. H. Kim, J. H. Na, H. Lim, J. H. Jeong, and K. Kim, "Fabrication and characteristics of novel load PMOS SSTFT (Stacked Single-crystal Thin Film Transistor) for 3-dimentional SRAM memory cell," in *Proc. IEEE Int. SOI Conf.*, 2004, pp. 127–129.

[17] Y.-S. Kwon, P. Lajevardi, F. Honor, and A. P. Chandrakasan, "A 3D FPGA wire resource prediction model validated using a 3D placement and routing rool," in *Proc. Syst. Level Interconnect Prediction Workshop (SLIP)*, 2005, pp. 65–72.

[18] K. W. Lee, T. Nakamura, T. Ono, Y. Yamada, T. Mizukusa, H. Hashimoto, K. T. Park, H. Kurino, and M. Koyanagi, "Three-dimensional shared memory fabricated using wafer stacking technology," in *Techn. Dig. Int. Electron Devices Meet.*, 2000, pp. 228–229.

[19] G. M. Link and N. Vijaykrishnan, "Thermal trends in emergent technologies," in *Proc. Int. Symp. Quality Electron. Des.*, 2006, pp. 625–632.

[20] G. Loh, Y. Xie, and B. Black, "Processor design in 3D die-stacking technologies," *IEEE Micro*, vol. 27, no. 3, pp. 31–48, May/Jun. 2007.

[21] M. Mamidipaka, K. Khouri, N. Dutt, and M. Abadir, "Analytical models for leakage power estimation of memory array structures," in *Proc. Int. Conf. Hardw./Softw. Codes. Syst. Synth. (CODES+ISSS)*, 2004, pp. 146–151.

[22] J. Mayega, O. Erdogan, P. M. Belemjian, K. Zhou, J. F. McDonald, and R. P. Kraft, "3D direct vertical interconnect microprocessors test vehicle," in *Proc. 13th ACM Great Lakes Symp. VLSI (GLSVLSI)*, 2003, pp. 141–146.

[23] J. Minz, E. Wong, and S. K. Lim, "Reliability-aware floorplanning for 3D circuits," in *Proc. IEEE Int. SOC Conf.*, 2005, pp. 61–62.

[24] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," presented at the 6th Int. Workshop Syst. Level Interconnect Prediction, Paris, France, 2004.

[25] V. Nguyen and P. Christie, "The impact of interstratal interconnect density on the performance of 3D ICs," in *Proc. Syst. Level Interconnect Prediction Workshop (SLIP)*, 2005, pp. 73–78.

[26] K. Puttaswamy and G. Loh, "Implementing caches in a 3D technology for high performance processors," in *Proc. Int. Conf. Comput. Des. (ICCD)*, 2005, pp. 525–532.

[27] K. Puttaswamy and G. Loh, "Dynamic instruction schedulers in a 3-dimensional integration technology," in *Proc. Great Lakes Symp. VLSI (GLSVLSI)*, 2006, pp. 153–158.

[28] K. Puttaswamy and G. Loh, "The impact of 3-dimensional integration on the design of arithmetic units," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2006, pp. 4951–4954.

[29] K. Puttaswamy and G. Loh, "Implementing register files for high-performance microprocessors in a die-stacked (3D) technology," in *Proc. IEEE Int. Symp. VLSI (ISVLSI)*, 2006, pp. 384–389.

[30] K. Puttaswamy and G. Loh, "Thermal analysis of a 3d die-stacked high-performance microprocessor," in *Proc. Great Lakes Symp. VLSI (GLSVSI)*, 2006, pp. 19–24.

[31] J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 2003.

[32] A. Rahman, S. Das, R. Reif, and A. P. Chandrakasan, "Wiring requirement and 3-D integration technology for FPGA," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 1, pp. 44–54, Jan. 2003.

[33] A. Rahman, A. Fan, and R. Reif, "Thermal analysis of 3D ICs," in *Proc. Int. Testing Conf. (ITC)*, 2001, pp. 157–159.

[34] A. Rahman and R. Reif, "System-level performance evaluation of three-dimensional integrated circuits," *IEEE Trans. Very Large Integr. (VLSI) Syst.*, vol. 8, no. 6, pp. 671–678, Jun. 2005.

[35] P. Shiu and S. K. Lim, "Multi-layer floorplanning for reliable system-on-package," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2004, pp. 69–72.

[36] P. Shivakumar and N. Jouppi, "Cacti 3.0: An integrated cache timing, power, and area model," Western Research Lab, Tempe, AZ, Res. Rep. 2001/2, 2001.

[37] Y.-F. Tsai, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Three-dimensional cache design exploration using 3D cacti," in *Proc. Int. Conf. Comput. Des.*, 2005, pp. 519–524.

[38] B. Vaidyanathan, W. Hung, F. Wang, Y. Xie, V. Narayanan, and M. J. Irwin, "Architecting microprocessor components in 3D design space," in *Proc. Int. Conf. VLSI Des.*, 2007, pp. 103–108.

[39] E. Wong and S. K. Lim, "3D floorplanning with thermal vias," in *Proc. Des., Autom. Test Euro. (DATE)*, 2006, pp. 878–883.

[40] A. Young, "Perspectives on 3D-IC technology," presented at the 2nd Annu. Conf. 3D Arch. Semiconductor Integr. Packag., San Francisco, CA, Jun. 2005.

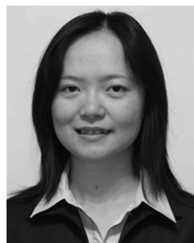
[41] A. Zeng, J. Lu, K. Rose, and R. J. Gutmann, "First-order performance prediction of cache memory with wafer-level 3D integration," *IEEE Des. Test Comput.*, vol. 22, no. 6, pp. 548–555, Jun. 2005.

[42] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, B. Zheng, Y. Wang, and M. Bohr, "A SRAM design on 65 nm CMOS technology with integrated leakage reduction scheme," in *VLSI Technol. Dig. Techn. Papers*, 2004, pp. 294–295.

[43] T. Zhang and S. Sapatnekar, "Thermal via placement in 3D Ics," in *Proc. ACM Int. Symp. Phys. Des. (ISPD)*, 2005, pp. 167–174.

[44] T. Zhang and S. Sapatnekar, "Temperature-aware routing in 3d ICs," in *Proc. Asia South Pacific Des. Autom. Conf. (ASPDAC)*, 2006, pp. 309–314.

[45] B. Black *et al.*, "Die stacking (3D) microarchitecture," in *Proc. Int. Symp. Microarch.*, 2006, pp. 469–479.



Yuh-Fang Tsai (S'00–M'05) received the B.S. degree in electronics engineering from Chun-Yuan Christine University, Chun-Li, Taiwan, R.O.C., in 1996, and the M.S. and Ph.D. degrees in computer science and engineering from the Pennsylvania State University, University Park, in 2002 and 2005, respectively.

Her research interests include power management and power aware VLSI circuit and system designs.

Feng Wang (S'05) received the B.S. degree in electronic engineering from Fundan University, Shanghai, China, in 1997, and the M.S. degree in electrical engineering from Ohio University, Athens, OH, in 2004. He is currently pursuing the Ph.D. degree in computer engineering from Pennsylvania State University, University Park.

His research interests include VLSI design, electronics design automation, and computer architecture.

Yuan Xie (SM'07) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, in 1999 and 2002, respectively.

He is an Assistant Professor with the Computer Science and Engineering Department, The Pennsylvania State University (Penn State), University Park. Before joining Penn State in the Fall of 2003, he was with IBM Microelectronic Division's Worldwide Design Center. His research interests include VLSI Design, computer architecture, embedded systems design, and electronics design automation.

Dr. Xie was a recipient of the SRC Inventor Recognition Award in 2002 and the National Science Foundation CAREER Award in 2006. He is a member of the ACM.

Narayanan Vijaykrishnan received the B.E. degree in computer science and engineering from University of Madras, Chennai, India, in 1993, and the Ph.D. degree in computer science and engineering from University of South Florida, Tampa, in 1998.

Since 1998, he has been with the Computer Science and Engineering Department, Pennsylvania State University, University Park, where he is currently an Associate Professor. His research interests include the areas of energy-aware reliable systems, embedded systems, nano/VLSI systems, and computer architecture.

Mary Jane Irwin (F'94) received the M.S. and Ph.D. degrees in computer science from the University of Illinois, Urbana-Champaign, in 1975 and 1977, respectively.

She is an Evan Pugh Professor and the A. Robert Noll of Engineering with the Department of Computer Science and Engineering, Pennsylvania State University, University Park. Her research and teaching interests include computer architecture (power constrained, application specific) and computer arithmetic, reliable systems design, and VLSI systems design and design automation.

Prof. Irwin was a recipient of an Honorary Doctorate from Chalmers University, Sweden, in 1997. She is an ACM Fellow and was elected to the National Academy of Engineering in 2003. She is currently serving as the Co-Chair of the ACM Publications Board and a member of the NRC Board of Army Science and Technology. In the past, she has served as an elected member of the IEEE Computer Societies Board of Governors, as Vice President of ACMs council, and one of Board of Directors of the Computing Research Association. She was the Editor-in-Chief of *ACM Transactions on the Design Automation of Electronic Systems* from 1998 to 2004 and Co-Editor-in-Chief of the *ACM Journal of Emerging Technologies in Computing Systems* from 2005 to 2006. She was the General Chair of the 1996 Federated Computing Research Conference, the 36th Design Automation Conference, the 2002 International Symposium on Low Power Electronics and Design, and the 2004 International Conference on Compilers, Architectures, and Synthesis for Embedded Systems.