Strategies for Improving the Parametric Yield and Profits of 3D ICs

Cesare Ferri
Division of Engineering
Brown University
Providence, RI 02912
cesare_ferri@brown.edu

Sherief Reda
Division of Engineering
Brown University
Providence, RI 02912
sherief_reda@brown.edu

R. Iris Bahar
Division of Engineering
Brown University
Providence, RI 02912
iris_bahar@brown.edu

ABSTRACT

Three-Dimensional (3D) Integrated Circuits (ICs) that integrate die with Through-Silicon Vias (TSVs) promise to continue system and functionality scaling beyond the traditional geometric 2D device scaling. 3D integration also improves the performance of ICs by reducing the communication time between different chip components through the use of short TSV-based vertical wires. This reduction is particularly attractive in processors where it is desirable to reduce the access time between the main logic die and the L2 cache or the main memory die. Process variations in 2D ICs lead to a drop in parametric yield (as measured by speed, leakage and sales profits), which forces manufacturers to speed bin their chips and to sell slow chips at reduced prices. In this paper we develop a model to quantify the impact of process variations on the parametric yield of 3D ICs, and then we propose a number of integration strategies that use a graph-theoretic framework to maximize the performance, parametric yield and profits of 3D ICs. Comparing our proposed strategies to current yield-oblivious methods, it is demonstrated that it is possible to increase the number of 3D ICs in the fastest speed bins by almost 2×, while simultaneously reducing the number of slow ICs by 29.4%. This leads to an improvement in performance by up to 6.45% and an increase of about 12.48% in total sales revenue using up-to-date market price models.

1. INTRODUCTION AND MOTIVATION

The overriding goals of Integrated Circuit (IC) technology is the development of circuits with increased functionality and performance at reduced power and cost. Historically, these increases have been achieved through 2D geometric scaling of devices and interconnects. Recently, three-dimensional (3D) ICs have emerged as an alternative or a complementary way to further continue or even accelerate the historic trend of increased functionality and performance. 3D ICs allow designers to stack die or wafers vertically in the same package, while connecting the communicating components on the different die using short Through-Silicon Vias (TSVs) [15, 33, 6]. Stacking multiple die in the same IC leads to a number of advantages, including (1) increased functionality in a smaller area, (2) improved performance through the use of shorter TSV-based vertical interconnects, and (3) reduced costs.

There are three main methods to achieve wafer-scale 3D integration [33]: *wafer-to-wafer, die-to-wafer, and die-to-die integration*. Wafer-to-wafer integration directly *bonds* entire wafers together. This method of integration has the advantage of high-throughput production, but it can lead to serious deterioration in fabrication yield as it can attach a good die to a faulty die, rendering the entire 3D IC faulty [33, 15]. Die-to-wafer uses a substrate wafer to integrate an already diced die on top of it. This offers high-throughput production together with high yield as it is possible to use only the known good die. Die-to-die integration allows the same high yield but suffers from low-production throughput. Die-to-wafer integration technology is the most likely candidate for future production-scale 3D ICs.

While the strategy for managing the *functional yield* in 3D integration looks simple: discard the faulty die and then just integrate the good with the good, the strategy for improving the *parametric yield*, as measured by speed, leakage and ultimately sales revenues, is not straightforward. An oblivious integration to process variations can determinately reduce the speed and parametric yield of 3D ICs by blindly integrating different die together.

A number of recent research efforts (e.g. [5, 17, 10, 24, 19]) focus on the developing design methodologies for 3D ICs. This paper focuses on the important issue of yield improvement in 3D ICs. The contributions of this paper are as follows.

1. This work is the first to examine the impact of process variation on the performance and parametric yield of 3D ICs. We formulate the general problem of optimizing the parametric yield in 3D integration under the presence of process variations.
2. Using a 3D processor as a 3D IC example, we model the impact of process variations on both the CPU and L2 cache die, and then we model the outcome of 3D integration on the performance of 3D processors.
3. This work is first to propose 3D integration strategies that exploit the flexibility of die-to-wafer and die-to-die integration to maximize the performance and parametric yield of 3D ICs.
4. Our strategies increase the number of 3D processors in the fastest bins by almost 2×, while simultaneously reducing the number of slow processors by 29.4% in comparison to current integration techniques. Our strategy also leads to an improvement in 3D processor performance (as measured by MIPS) by up to 6.45% and an increase of about 12.48% in total sales revenue using up-to-date market price models.

The organization of this paper is as follows. Section 2 gives an overview of the necessary background for this work. In Section 3, we (1) formulate the problem of improving the parametric yield in 3D ICs; (2) show how to model the performance of 3D CPU-memory stacks under the presence of process variations; and (3) propose a number of strategies to optimize the parametric yield. Section 4 gives an extensive set of experimental results supporting our methodology, and finally Section 5 summarizes the main results of this paper.

1-4244-1382-6/07/$25.00 ©2007 IEEE

2. BACKGROUND

This paper examines the problem of improving yields, in particular parametric yield, during the fabrication of 3D ICs. *Yield* is defined as the number of functionally good dies from the total manufactured ones. *Parametric yield* is concerned with the number of functional dies that meet the required performance and power constraints set by the chip designers. In this section, we first briefly describe the impact of 3D systems on IC architecture and design methodologies, and then we discuss how process variations change the electrical properties of ICs and its impact on parametric yield.

Benefits and Challenges of 3D Integration. 3D ICs allow the creation of new systems that are currently not feasible by planar fabrication technology. Using a 3D approach allows the integration of dissimilar technologies to create hybrid chips that include memory, logic, optical, RF, and analog components. As mentioned in the introduction, there are a number of different fabrication techniques that can produce 3D chips [33, 15, 6].

Besides improved functionality and system scaling capabilities, 3D integration also promises to replace long 2-D interconnects by short TSV-based vertical interconnects [15, 33]. Long (or global) 2-D interconnects have large delays [14] and require an increasing number of repeaters to appropriately buffer them [32]. By transforming long 2-D interconnects into short 3-D vertical interconnects with less capacitive and resistive loading, the system delay is improved [33]. Reducing long interconnect delay is especially important for processors, as they continuously access memory subsystems. With 3D integration, processors can cut down the memory access time which improves the overall system performance. The quantification of this improvement has been the subject of a number of recent works [35, 25, 21]. For example, Zeng *et al.* [35] compare the performance of cache design in 2D and 3D ICs and develop a predictor tool for 3D CPU/cache stack access time prediction. The tool is similar to a 2D cache access time predictor (e.g. [34]), but takes vertical interconnects into account. While increased functionality and performance are the major advantages of 3D integration, 3D integration also brings its own challenges in terms of fabrication, production yield and heat removal from the stacked chips [15, 24]. In this paper, we address 3D IC performance improvement within the context of process variations and parametric yield improvement.

Process variations and Yield. The impact of process variations on circuit and architectural performance of 2D ICs has been the subject of recent investigations [7, 26, 8, 20, 18]. Process variations change the electrical parameters of ICs from their designers original intent. Process variations can heavily impact the frequency and leakage power of CPU cores [8, 7, 26, 20, 22] as well as the access time and leakage of memory subsystems [18, 27]. Fabrication facilities typically categorize chips according to their performance by *speed binning* them and assigning them to appropriate price points [12, 13]. Improving the parametric yield is concerned with optimizing the values of the electrical parameters of fabricated chips in order to achieve an overall good performance and profits [31, 13].

Yield of 3D ICs. Yield loss is considered as one of the bottlenecks that need to be overcome to bring 3D technology from the lab to the fab and the marketplace [4]. Despite its importance, the problem of yield improvement of 3D ICs has been least investigated in the literature. A number of recent efforts [5, 33, 30] point to the importance of functional yield management of 3D ICs. This work is the first to investigate the problem of process variation modeling and parametric yield improvement in 3D ICs.

3. PROPOSED YIELD IMPROVEMENT METHODOLOGY

The objective of this section is to formulate the general problem of optimizing the parametric yield of 3D integration under the presence of process variability (Subsection 3.1) and then propose solutions to achieve optimal yield. Quantifying or modeling the impact of process variations on the parametric yield of 3D ICs is more complex than in 2D ICs as different die that belong to the same 3D IC are fabricated on separate wafers and then integrated and interconnected.

3.1 Problem Statement and Formulation

3D ICs can have two or more die stacked and integrated with TSVs. The general problem of optimizing the parametric yield of 3D ICs under process variations can be stated as follows.

Problem 1. Given K different wafers (or wafer lots) each with identically N designed dies, yet the dies are parametrically different due to process variations, find an integration assignment or matching strategy that maximizes the total parametric yield of the N produced 3D ICs, where each IC is composed of K stacked die.

The outline solution for this problem is as follows.

1. Model the impact of the process variations on both the speed and leakage on each die for all K wafers.
2. Model the performance of the 3D system (composed of K different dies) for every possible N^K 3D IC combination.
3. From the N^K possible combinations, find the N combinations that maximize the total parametric yield (as measured by performance, leakage or revenue) such that each die is assigned to exactly one 3D IC package.

The problem is obviously electrically and combinatorially challenging. First, the impact of process variations on the electrical properties (speed and leakage) have to be modeled for each die and for each possible 3D combination, and second, the N 3D IC combinations that maximize the total parametric yield have to be computed and selected. For the case of three or more integrated die, one can prove that maximizing the parametric yield for 3D ICs is NP-hard by reducing the classical NP-hard 3-D matching problem [16] to it. A more tractable version of the problem is possible in the case of two die, i.e., $K = 2$, where for example the first wafer holds processor logic and the second wafer holds the processor L2 cache (or memory in general).

The problem is illustrated by Figure 1. The upper wafer holds a set of die, say L2 cache or memory die. The die had been tested and the faulty have been identified (labeled with **F**) and the good have been labeled with their speed. The same testing and labeling procedure has been carried out for the die in the substrate Central Processing Unit (CPU) or logic wafer. Our problem seeks to find an integration strategy that maximizes the total parametric yield. To solve the problem, we first model the impact of process variations on the speed of both the CPU and L2 cache die (Subsection 3.2). Then we model the outcome of integrating a CPU and L2 cache on the performance of 3D processor. In Subsection 3.3, we propose a number of combinatorially-tractable strategies to optimize the total parametric yield. In Subsection 3.4, we extend our strategies to directly maximize 3D chip sales revenues.

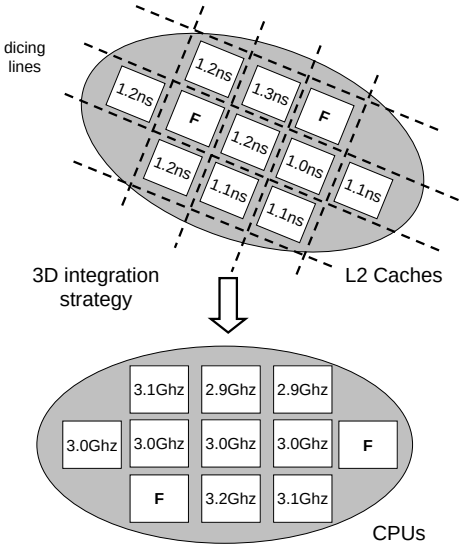


Figure 1: 3D integration strategies to maximize the parametric yield. F indicates a faulty die. The number inside each die represent its speed as measured by testing before 3D integration.

3.2 Modeling the impact of process variations

To model the impact of process variations on 3D ICs, it is necessary to first model the impact of process variation on the individual die and then model the interplay of the process variations on the different die composing a 3D IC.

In 2D ICs, process variations can be categorized in *intra-die variations*, which affect sub-parts of a single chip, and in *inter-die variations*, which affect the performance and power parameters of different chips [8]. The overall impact of intra and inter-die variations is that they lead to considerable discrepancies in the performance of fabricated chips. The distribution of chips as a function of performance typically exhibits a Gaussian-like form [29, 8], where the mean and the standard deviation of the distribution are functions of the *intra-die* and *inter-die* variations respectively.

Modeling the variations in CPU die. To model the impact of process variations on the CPU die, we first recall the fact that the speed of a CPU is determined by the clock period of its critical path. Following the modeling strategy of [8], we design a circuit with 9 levels of NAND gates to represent a CPU critical path. We model the process variations by generating a vector of random, Gaussian distributed numbers. Each number represents the nominal value of the gate length for an entire CPU. Then, SPICE [1] is used to simulate the critical path for each gate length value, and maximum possible clock frequencies (Ghz) results.

Modeling the variations in L2 die. The L2 cache modeling relies on the same statistical analysis principles. Contrary to our previous strategy, we do not simulate the L2 cache critical path directly, but rather, we use the PRACTICS tool [35] to first obtain the nominal delay values of L2 caches stacked in 3D chips. Then we impose a Gaussian distribution around the nominal values. More details can be found in Subsection 4.2.

Modeling the variations in 3D ICs. A 3D processor is the product of integrating two or more 2D die: one CPU die and one L2 cache die. Both the CPU and the L2 cache variations contribute to the variability for the entire 3D structure. These variations are going to ultimately change the performance of a 3D processor. To quantify this impact, we choose the popular *MIPS (millions of instructions per second)* as the performance index. For a given pair (i, j) of CPU i and L2 cache j, we compute the MIPS in the following way. We first calculate the L2 latency, $L_{i,j}$, in terms of CPU cycles, i.e., $L_{i,j} = \lceil \frac{\text{L2 } j \text{ access time}}{\text{CPU } i \text{ cycle period}} \rceil$. While the access time and core frequency may display wide variations, the ceiling rounding, by the $\lceil \cdot \rceil$ operator, reduces the number of distinct latency values. Note that the L2 latency varies from a minimum of $L_{\min} = \lceil \text{min L2 access time} \times \text{min CPU frequency} \rceil$ to a maximum of $L_{\max} = \lceil \text{max L2 access time} \times \text{max CPU frequency} \rceil$.

The impact of cache latency on performance depends on the particular application executing on the processor. A memory-intensive application will be heavily impacted by large values of latency in comparison to a processing-intensive application. To obtain an accurate estimation of the overall speed of the 3D processor, the cache latency values are fed to an architectural simulator (e.g. SimpleScalar [9]) to compute the actual *Cycles Per Instruction (CPI)* using typical benchmark applications. The fact that there are only a few possible values for the L2 latency drastically reduces the number of architectural simulations that need to be carried out. Finally, the MIPS of the 3D chip composed of CPU i and cache j is simply the clock frequency of i multiplied by the CPI of the pair.

3.3 Proposed 3D Integration Strategies

While wafer-to-wafer bonding dictates the outcome of integrating different wafers, die-to-wafer and die-to-die integration offer flexibility that we propose to exploit by devising 3D integration strategies that maximize the parametric yield. We propose a number of strategies that control and improve the parametric yield of 3D ICs. In this subsection, we focus on improving the parametric yield as measured by the speed or performance of the 3D package. The strategies vary in their ability to optimize the parametric yield, and also in their computational complexity.

- **Random-Random (RR) assignment.** In this naive strategy, the 3D integration process is oblivious or blind to parametric yield and assigns CPUs and L2 caches randomly to form the 3D processor chips. This strategy can be used as a baseline to compare other strategies against.

- **Fast-Fast (FF) assignment.** In this strategy, CPU die are sorted in descending order (fastest first) according to their tested speed (CPU frequency), and then L2 cache die are sorted in ascending order (fastest first) according to their tested speed (access time). Then the 3D chips are constructed by matching the CPUs and L2 caches in order. This strategy starts pairing the fastest CPUs and L2 caches together and ends pairing the slowest CPUs and caches together. This strategy attempts to obtain the fastest possible 3D processor chips (at the cost of producing the slowest possible 3D chips). This strategy is easily computed in $O(N \log N)$ runtime.

- **Fast-Slow (FS) assignment.** In this strategy, CPU die are sorted descending order (fastest first) according to their tested speed (CPU frequency), and then L2 caches are sorted ascending order (slowest first) according to their tested speed (access time). Then the 3D chips are constructed by matching the CPUs and L2 caches in order. This strategy starts pairing the fastest CPUs with the slowest L2 caches together

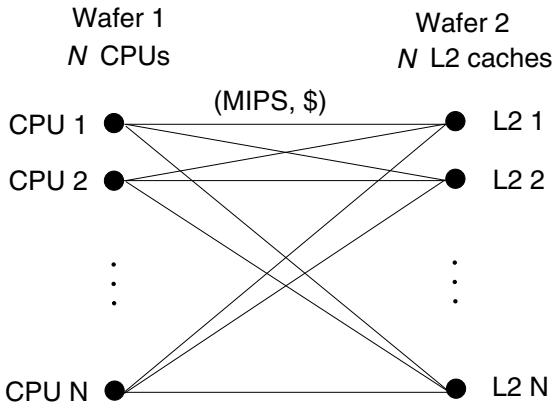


Figure 2: Optimal assignment of CPUs to L2 caches to generate 3D processor chips that maximize the total parametric yield as measured by the total system MIPs.

and ends pairing the slowest CPUs with the fastest cache together. This strategy attempts to increase the number of processors with medium speed. This strategy is easily computed in $O(N \log N)$ runtime.

- **Optimal (OPT) Assignment.** The yield maximization problem can be formulated using a graph-theoretic framework. In this case, *vertices* represent the die, *edges* represent the possible 3D ICs, and *edge costs* represent the yield (speed or revenue) value of the possible ICs. Thus, we construct a bipartite graph, given in Figure 2, with $2N$ vertices representing the N CPU die and the N L2 cache die, and N^2 edges where each edge is labeled by the MIPS of the 3D processor produced from the CPU and L2 cache die that are its end points. The optimal assignment strategy involves finding the N CPU/L2 pairs that maximize the total MIPS, and such that each CPU or L2 cache participate in only one 3D IC. The optimal assignment can be found by computing the maximum graph matching or assignment in the bipartite graph. This can be computed in polynomial $O(N^3)$ runtime using the classical Hungarian algorithm [23, 28].

The performance of a 3D system determines its speed bin and consequently its price. This is described in the next subsection.

3.4 Strategies to Maximize Sales Profits

Subsections 3.2 and 3.3 give us a way to model the MIPS of 3D processors under processor variations and guide their 3D integration process. Besides maximizing processor MIPS, chip manufacturers are ultimately interested in maximizing sales profits. Processors with higher performance (measured by MIPS) are naturally sold with higher prices than the lower ones. The difference in price is correlated with the number of available supply chips. As process variations produces chips with Gaussian-like distributions as we have explained earlier, it is expected that there are very few chips with extremely high or extremely low performance and the majority of chips have a performance around some average value. This leads to a non-linear relationship between the performance and price of the chip. For example, the market values of Intel Core Duo processors (according to pricegrabber.com) for its different four speed bins are given in Figure 3. The plot shows an exponential trend for the price. The price of extreme processors are almost double the price of the fast processors, which are in turn double the price of the slow ones.

Our proposed fast-fast and optimal-assignment strategies are designed to increase the number of fastest 3D chips (as we will confirm in Section 4). Thus it is likely that this leads to a significant increase in total sales profits according to the market price model. It is also possible to directly derive our optimal assignment strategy, described in Subsection 3.3, using the dollars values of the 3D system, rather than using the MIPS value. In that case, we can substitute the MIPS label of each edge in Figure 2 by the corresponding dollars value and find the optimal assignment strategy as described earlier.

4. EXPERIMENTAL RESULTS

In this section we quantify the impact of our 3D integration strategies on the parametric yield and profits of 3D ICs. To realistically achieve this goal, we design a tool chain flow, given in Figure 4, to calculate the performance (as measured by MIPS) and the parametric yield of 3D ICs. We use the following tools.

- SPICE to calculate the delay of CPUs under the presence of process variations at 70nm [2].
- CACTI (version 4.2) [34] and PRACTICS (version 1.0) [35] to calculate the access time of L2 caches using vertical interconnects in 3D chips.
- SimpleScalar (version 3.0) [9] to model the performance (as measured by Cycles Per Instruction CPI) of 3D processors given the underlying CPU frequency and the L2 cache access time with vertical interconnects.
- The matching code by Rohe [11] to compute the optimal 3D assignment strategy to maximize the parametric yield.

Out tool chain starts by modeling the impact of process variations on the speed of 100 CPU and the L2 cache die. Then SimpleScalar is used to calculate the performance of the potential 3D processors composed of the different CPU and L2 die. This information is used to construct a bipartite graph as outlined in Subsection 3.3 which is then fed, with market price models, to the optimal matching module to find the integration assignment that maximizes the parametric yield and profits. In the following subsections, we describe each tool and step in detail.

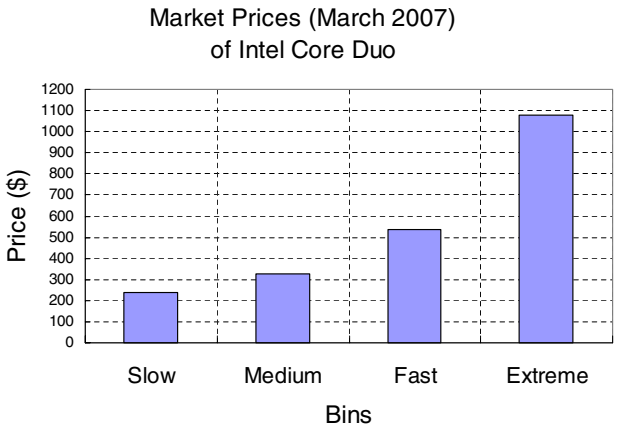


Figure 3: Market prices (according to pricegrabber.com of Intel Core Duo as of March 2007.

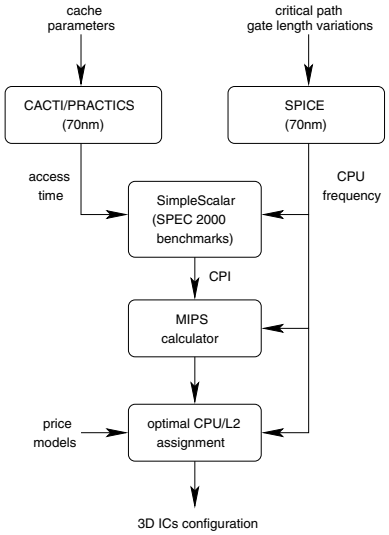
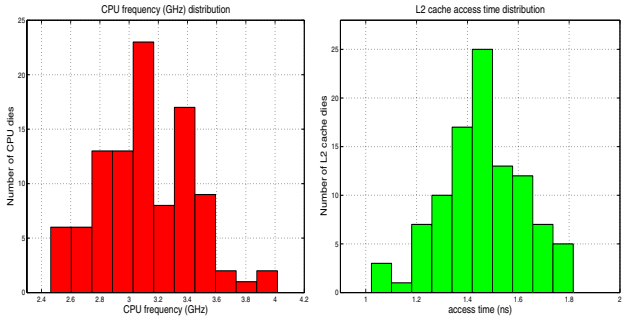


Figure 4: The tool chain required to model and evaluate our strategy.



(a) CPUs frequency(Ghz) distribution as produced by modeling critical path variations on the CPU delay using SPICE.

(b) L2 cache access time (ns) distribution as produced from modeling process variations using Cacti and PRACTICS.

Figure 5: Impact of process variations on CPU frequency and L2 cache access time.

4.1 CPU Setup

As described in Section 3.2, we quantify the impact of process variations on the performance and power of CPUs by simulating with SPICE a typical CPU critical path (i.e. a chain of 9 NAND gates representing the CPU pipeline flow [8]). We use the 70nm Berkeley predictive technology model for all the simulations [2]. To model the impact of inter-die process variations, we generate 100 critical path SPICE netlists, where the gate length of each is drawn from a Gaussian distribution with a mean of 70nm and standard deviation of 5.07% (leading to a $\pm$10nm maximum variations). We then execute SPICE on each netlist and record the delay and leakage current consumed. The frequency of a CPU is the reciprocal of the critical path delay. We plot the distribution of CPU frequencies (GHz) obtained from SPICE simulations in Figure 5(a). Table 1 gives the maximum, minimum, average and standard deviation of the distribution of CPU frequencies. The distribution of CPUs have a standard deviation of 10.33% with a mean of 3.12GHz.

parameter	CPU		parameter	L2 cache	
frequency	max	4.01GHz	access time	max	1.81ns
	average	3.12GHz		average	1.46ns
	min	2.46GHz		min	1.02ns
	std dev.	10.33%		std dev.	11.06%

Table 1: Impact of process variations on the speed of CPU and L2 cache dies.

4.2 L2 Cache Setup

Assuming a cache configuration of 2MB at the 70nm technology node, we calculate the cache access time using PRACTICS [35], which is a tool for predicting the access time of L2 caches using vertical interconnects in 3D ICs. After calculating the access time, we generate 100 random Gaussian distributed access times with an average equal to that reported by the PRACTICS tool and an imposed standard deviation equal to 11.06%. Figure 5.b plots the resulting distribution for the L2 cache access time (ns). The main statistics of the L2 access time distribution are reported in Table 1.

4.3 3D Processor Performance Modeling

With the computed CPU frequency vector (100 values in GHz) and the L2 access time vector (100 values in ns), we calculate the L2 access time in terms of CPU cycles, $\lceil \frac{\text{L2 access time}}{\text{CPU cycle period}} \rceil$, for every possible pair of CPU and L2 cache. While the number of different CPU frequencies and cache access times could be large due to process variations, the number of distinct different cache access cycles are much fewer in number (e.g. they vary between 3 to 8 cycles). The newly computed values for access cycles are used as configuration parameters for the SimpleScalar simulator (which requires cache access time expressed in CPU cycles) to simulate the performance of every possible CPU/L2 3D chip combination[1]. Next, we run a suite of six SPEC 2000 (three integer and three floating-number based) benchmarks [3] and compute the average Cycles Per Instruction (CPI) over the six benchmarks: three integer benchmarks (gcc, parser, gzip), and three floating point applications (mgris, apsi, equake). CPI results are given in Table 2. We then use the CPI and clock frequency values to calculate the MIPS of every possible CPU/L2 3D processor.

4.4 Evaluation of 3D Integration Strategies

With the modeled CPU frequency, L2 access time, and processor MIPS, it is possible to evaluate the effectiveness of our different 3D integration strategies on the parametric yield as measured by the performance of the 3D processor. Given the CPU frequency and L2 access time distributions of Figure 5, we compute the MIPS distributions of 3D processors produced by different assignment strategies (RR, FF, FS and OPT). We report the performance in terms of average, maximum, and minimum MIPS in Table 3. The results of Table 3 demonstrate that the optimal assignment strategy and the fast-fast strategy produce systems with the maximum MIPS; however, the optimal strategy has the highest average MIPS of all strategies. Compared to the performance oblivious strategy (the random-random strategy), the optimal assignment strategy produces system with better performance by up to 6.49% with an average improvement of 1.71%.

The processor distributions produced from different integration

[1]We use the following parameters for simulation: (1) 2-way, 3 cycle L1 cache of 16 Kbyte; (2) 8-way 2MB L2 cache; (3) main memory latency is 50 cycles; (4) the decode/issue/commit width is 4 issue.

L2 Latency (cycles)	bench	CPI	Avg CPI	L2 Latency (cycles)	bench	CPI	Avg CPI
3	apsi	0.614	0.734	6	apsi	0.614	0.847
	equake	0.785			equake	0.905	
	gcc	1.031			gcc	1.556	
	gzip	0.577			gzip	0.594	
	mgrid	0.548			mgrid	0.547	
	parser	0.850			parser	0.863	
4	apsi	0.614	0.798	7	apsi	0.615	0.873
	equake	0.865			equake	0.927	
	gcc	1.330			gcc	1.669	
	gzip	0.585			gzip	0.599	
	mgrid	0.543			mgrid	0.559	
	parser	0.851			parser	0.868	
5	apsi	0.614	0.819	8	apsi	0.615	0.899
	equake	0.875			equake	0.955	
	gcc	1.441			gcc	1.786	
	gzip	0.585			gzip	0.604	
	mgrid	0.546			mgrid	0.561	
	parser	0.855			parser	0.876	

Table 2: CPI reported for different L2 cache access cycles ($\lceil \frac{\text{L2 access time}}{\text{CPU cycle period}} \rceil$). L2 access times and CPU clock periods are taken from the data of Figure 5.

Strategy	Max MIPS	Average	Min MIPS	Δ MIPS (%)
Fast-Fast	4902.62	3810.41	3006.78	63.04%
Fast-Slow	4465.68	3784.63	3221.22	38.63%
Optimal	4903.00	3855.00	3138.00	56.25%
Random-Random	4606.61	3790.17	3082.71	49.93%

Table 3: Impact of different 3D integration strategies on the statistical performance parameters of 3D processors chips. We calculate Δ MIPS (%)= $\frac{\text{Max MIPS}-\text{Min MIPS}}{\text{Min MIPS}}$.

strategies are also given in Figure 6. To create the figure, we use the RR processor distribution to designate four performance bins: extreme, fast, medium, and slow, using Matlab's histogram function. The four bins mimic those of Intel processors as we described earlier in Subsection 3.4. We use the bin boundaries of the RR strategy as the bin boundaries of other 3D integration strategies. This way we guarantee a fair comparison for the different strategies. From the data, we draw the following observations.

- The OPT and FF strategies produce almost twice the number of extreme processors compared to other strategies.
- While OPT and FF produce the same number of Extreme processors, OPT reduces the number of processors in the slow bin by almost half compared to FF. Note that FF produces the highest number of CPUs in the slow bin.
- FS produces a large number of CPUs in the medium and fast bins, but produces the fewest number of CPUs in the extreme bin.

4.5 Impact on Revenue

As we discussed earlier in Subsection 3.4, after fabrication and binning, IC manufacturers price according to their bin. We also follow the same strategy with the 3D chips produced from our different integration strategies. With the number of processors in each bin in hand from Figure 6, we readily calculate the total revenues from applying different integration strategies and report them in Table 4. We multiply the number of processors in each bin by the

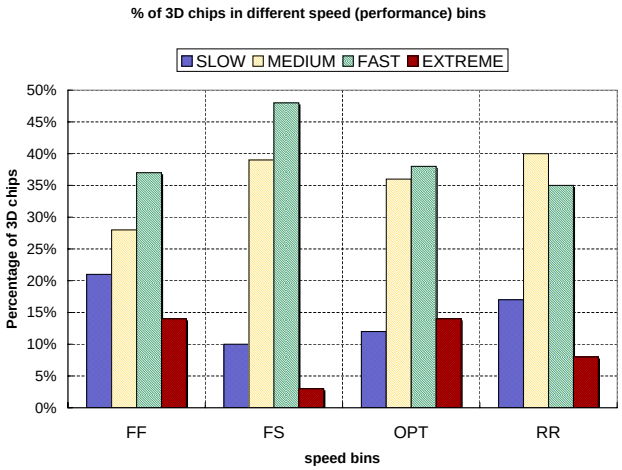


Figure 6: Impact of process variations on 3D processor performance as measured by MIPS using the different proposed 3D integration assignment strategies. Four performance or speed bins are used: slow, medium, fast, and extreme.

market price of the bin (as given in Figure 3 according to pricegrabber.com) and sum over all bins to give the total revenues. The results show that the optimal strategy yields an increase of 12.48% in total revenues compared to the random-random strategy. The FF strategy comes second with an increase of 10.28%. As long as market prices favor systems with extreme performance, FS can lead to a loss in revenues.

4.6 Practical Considerations at Fabrication

While we have resorted to modeling and simulations to evaluate the impact of process variations on the CPU and L2 cache dies, the situation is actually simpler at a fabrication facility. Speed testing is routinely done to label the speed of each die. These speeds are then used to calculate the access time of each L2 cache die in terms of CPU cycles for each possible CPU/L2 combinations. The handful of values obtained for cache latency in cycles are used as indices to a pre-computed CPI lookup table (e.g. Table 2). The only thing that needs to be computed during fabrication runtime is the assignment algorithm to figure out the optimal way to integrate the various CPUs and die.

5. CONCLUSIONS

In this paper, we have investigated the problem of maximizing performance, parametric yield and profits in 3D integrated circuits. We have described how to model the impact of process variations on 3D ICs and we have proposed a number of integration strategies that maximize the total parametric yield and profits. Using a 3D processor as an example, we have demonstrated that our optimal assignment scheme leads to an overall 6.5% improvement in performance and 12.5% increase in revenue. In a market where profit margins for computer systems may be relatively small, this increase in revenue can translate to a substantial increase in profits. Comparing the greedy fast-fast strategy to the optimal strategy, we find that optimal matching strategy simultaneously reduces the total number of slowest processors almost in half and maximizes the number of fastest processors. Our near-future work will incorporate leakage into the proposed integration strategies. In particular we will develop strategies that constrain the maximum leakage in 3D ICs.

Bin	Market price ($) per chip	3D Integration Strategy							
		Fast-Fast		Fast-Slow		Optimal		Random-Random	
		#chips (%)	Revenues ($)	#chips (%)	Revenues ($)	#chips (%)	Revenues ($)	#chips (%)	Revenues ($)
EXTREME	1081	14	15134	3	3243	14	15134	8	8684
FAST	538	37	19906	48	25824	38	20444	35	18830
MEDIUM	325	28	9100	39	12675	36	11700	40	13000
SLOW	240	21	5040	10	2400	12	2880	17	4080
Total		100	49180 (10.28%)	100	44142 (-1.01%)	100	50158 (12.48%)	100	44594 (0.00%)

Table 4: Impact of different 3D assignment strategies on the number of processor in each speed bin as well as the total revenue.

6. REFERENCES

[1] [Online]. Available: www.simucad.com

[2] [Online]. Available: http://www.eas.asu.edu/ ptm/introduction.html

[3] [Online]. Available: http://www.spec.org/cpu/

[4] J. Baliga, "Chips Go Vertical," *IEEE Spectrum Magazine*, vol. 41(3), pp. 43–47, 2004.

[5] K. Banerjee, S. J. Souri, P. Kaput, and K. C. Saraswat, "3-D ICs: A Novel Chip Design for Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration," *Proceedings of the IEEE*, vol. 89(5), pp. 602–633, 2001.

[6] P. Benkart, A. Kaiser, A. Munding, M. Bschorr, H.-J. Pfleiderer, E. Kohn, A. Heittmann, H. Huebner, and U. Ramacher, "3D Chip Stack Technology Using Through-Chip Interconnects," *IEEE Design & Test of Computers*, vol. 22(6), pp. 512–518, 2005.

[7] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter Variations and Impact on Circuits and Microarchitecture," in *Design Automation Conference*, 2003, pp. 338–342.

[8] K. Bowman, S. Duvall, and J. Meindl, "Impact of Die-to-Die and Within-Die Parameter Fluctations on the Maximum Clock Frequency Distribution for Gigascale Integration," *IEEE Journal of Solid State Electronics*, vol. 37(2), pp. 183–190, 2002.

[9] D. C. Burger and T. M. Austin, "The SimpleScalar Tool Set, Version 2.0, Tech. Rep. CS-TR-1997-1342, 1997. [Online]. Available: citeseer.ist.psu.edu/burger97simplescalar.html

[10] J. Cong, J. Wei, and Y. Zhang, "A Thermal-Driven Floorplanning Algorithm for 3D ICs," in *Proc. Intl. Conference on Computer Aided Design*, 2004, pp. 306–313.

[11] W. Cook and A. Rohe, "Computing Minimum Weight Perfect Matchings, http://www.or.uni-bonn.de/home/rohe/matching.html," *INFORMS J. Computing*, vol. 11, pp. 38–148, 1999.

[12] B. D. Cory, R. Kapur, and B. Underwood, "Speed Binning with Path Delay Test in 150-nm Technology," *IEEE Design & Test of Computers*, vol. 20(5), pp. 41–45, 2003.

[13] A. Datta, S. Bhunia, J. H. Choi, S. Mukhopadhyay, and K. Roy, "Speed Binning Aware Design Methodology to Improve Profit under Parameter Variations," in *Proc. Asia and South Pacific Design Automation Conference*, 2006, pp. 712–717.

[14] J. A. Davis, R. Venkatesan, A. Kaloyeros, M. Beylansky, S. J. Souri, K. Banerjee, K. C. Saraswat, A. Rahman, R. Reif, and J. Meindl, "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century," *Proceedings of the IEEE*, vol. 89(3), pp. 305–324, 2001.

[15] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. Sule, M. Steer, and P.D.Franzon, "Demystifying 3D ICs: The Pros and Cons of Going Vertical," *IEEE Design & Test of Computers*, vol. 22(6), pp. 498–510, 2005.

[16] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, first (twenty-third printing) ed. W.H. Freeman and Company, 1979.

[17] B. Goplen and S. Sapatnekar, "Thermal Via Placement in 3D ICs," in *Proc. Intl. Symposium on Physical Design*, 2005, pp. 167–174.

[18] E. Grossar, M. Stucchi, K. Maes, and W. Dehaene, "Statistically Aware SRAM Memroy Array Design," in *International Symposium on Quality Electronic Design*, 2006, pp. 25–30.

[19] M. Healy, M. Vittes, M. Ekpanyapong, C. S. Ballapuram, K. S. Lim, H.-H. S. Lee, and G. H. Loh, "Multiobjective Microarchitectural Floorplanning for 2-D and 3-D ICs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26(1), pp. 38–52, 2007.

[20] E. Humenay, D. Arjan, and K. Skadron, "Impact of Parameter Variations on Multi-Core Chips," in *Workshop on Architectural Support for Gigascale Integration*, 2006, pp. 1–9.

[21] P. Jacob, O. Erdogan, A. Zia, P. M. Belemjian, R. P. Kraft, and J. F. McDonald, "Predicting the Performance of a 3D Processor-Memory Chip Stack," *IEEE Design & Test of Computers*, vol. 22(6), pp. 540–547, 2005.

[22] C. Kim, J.-J. Kim, I.-J. Chang, and K. Roy, "PVT-Aware Leakage Reduction for On-Die Caches with Improved Read Stability," *IEEE Journal of Solid-State Circuits*, vol. 41(1), pp. 170–178, 2006.

[23] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistic Quarterly*, vol. 2, pp. 83–97, 1955.

[24] S. K. Lim, "Physical Design for 3D System on Package," *IEEE Design & Test of Computers*, vol. 22(6), pp. 532–539, 2005.

[25] C. C. Liu, I. Ganusov, M. Burtscher, and S. Tiwari, "Bridging the Processor-Memory Performance Gap with 3D IC Technology," *IEEE Design & Test of Computers*, vol. 22(6), pp. 556–564, 2005.

[26] D. Marculescu and E. Talpes, "Variability and Energy Awareness: A Microarchitecture-Level Perspective," in *Design Automation Conference*, 2005, pp. 11–16.

[27] K. Meng and R. Joseph, "Process Variation Aware Cache Leakage Management," in *International Symposium on Low-Power Electronics*, 2006, pp. 262–267.

[28] J. Munkres, "Algorithms for the Assignment and Transportation Problems," *Journal of the Society of Industrial and Applied Mathematics*, vol. 5(1), pp. 32–38, 1957.

[29] M. Orshansky, L. Milnor, P. Chen, K. Keutzer, and C. Hu, "Impact of Spatial Intrachip Gate Length Variability on the Performance of High-Speed Digital Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21(5), pp. 544–553, 2002.

[30] R. S. Patti, "Three-Dimensional Integrated Circuits and the Future of Systems-on-Chip Designs," *Proceedings of IEEE*, vol. 94(6), pp. 1214–1224, 2006.

[31] R. R. Rao, D. Blaauw, D. Sylvester, and A. Devgan, "Modeling and Anlysis of Parametric Yield Under Power and Performance Constraints," *IEEE Design & Test of Computers*, vol. 22(4), pp. 376–385, 2005.

[32] P. Saxena, N. Menezes, P. Cocchini, and D. A. Kirkpatrick, "Repeater Scaling and Its Impact on CAD," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 23(4), pp. 451–463, 2004.

[33] A. W. Topol, J. D. C. La Tulipe, L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, and M. Ieong, "Three-dimensional Integrated Circuits," *IBM Journal of Res. and Dev.*, vol. 50(4-5), pp. 491–506, 2006.

[34] S. Wilton and N. P. Jouppi, "CACTI: An Enhanced Cache Access and Cycle Time Model," *IEEE Journal Solid-State Circuits*, vol. 31(5), pp. 677–688, 1996.

[35] A. Zeng, J. Li, K. Rose, and R. J. Gutmann, "First-Order Performance Prediction of Cache Memory with Wafer-Level 3D Integration," *IEEE Design & Test of Computers*, vol. 22(6), pp. 548–555, 2005.