# 3D Architecture Modeling and Exploration

Jason Cong[†] Eren Kursun[†] Yongxiang Liu[†] Yuchun Ma[‡] Glenn Reinman[†]

[†]Computer Science Department, University of California, Los Angeles
[‡] Tsinghua University, China

**Abstract**

*Vertical integration (3D ICs) has demonstrated the potential to reduce inter-block wire latency through flexible block placement and routing. However, there is untapped potential for 3D ICs to reduce intra-block wire latency through architectural designs that can leverage multiple silicon layers in innovative ways. Furthermore, it is particularly challenging to simultaneously explore the physical design space and microarchitectural space for vertical integration. The physical design space typically has no information on the microarchitectural impact of latency optimization, and the microarchitectural space has no information on the physical design impact of different architectural alternatives.*

*We make the following contributions in this paper: (1) the introduction of port partitioning, a new approach to constructing multi-layer blocks, (2) the extension of a microarchitectural exploration tool to include the ability to model multi-layer blocks and to consider these blocks as alternative implementations of single layer architectural blocks on the fly, within a single floorplanning run, and (3) the evaluation of vertical integration on a design driver using this framework.*

*For this design driver, we see an average 36% improvement in performance (measured in BIPS) over a single layer architecture, and a 29% improvement in performance over a multi-layer architecture with single layer blocks. The on-chip temperature is kept below* $40^{\circ}\mathrm{C}$.

## 1 Introduction and Motivation

Vertical integration [20, 27, 32, 29, 12] leverages multiple layers of silicon to allow physical designers more flexibility in component layout. One approach to using this technology is to place single-layer (i.e. 2D) blocks in one of the silicon layers and running both horizontal and vertical interconnect between blocks. The flexibility that this design affords has the potential to dramatically reduce inter-block interconnect latency in a design [7, 1, 2, 6].

However, this approach does little to help intra-block wire latency. And despite the advantage of almost completely eliminating inter-block wire latency, we find that the placement of 2D blocks in two layers improves performance by 6% on average for a particular architecture (described in section 5). Additional gains from the use of vertical integration must attack the intra-block wire latency.

Furthermore, the emergence of technology like vertical integration can have a dramatic impact on microarchitecture design – a field that is heavily reliant on physical planning and technological innovation. However, physical planning is not meaningful without consideration for microarchitectural loop sensitivities: some loose loops [3] are better able to tolerate latency than others [28]. A floorplan with a 5% reduction in wirelength may actually be better than a floorplan with a 7% reduction in wirelength – if the former reduces the length of more critical microarchitectural loops than the latter. Similarly, architectural innovations are not meaningful without understanding their physical design implications.

Recently, the MEVA-3D [6] framework was proposed to bridge the gap between physical planning and microarchitectural design. The framework uses microarchitectural loop sensitivities in the floorplanning process to guide block placement. With this framework, architects can obtain accurate loop latencies to feed to a cycle-accurate simulation framework. This can help evaluate the impact of new and emerging technologies on microprocessor design.

In this paper, we explore the architectural impact and potential of finer granularity vertical integration, where individual blocks are placed across multiple layers. The challenge from the architectural side is the construction of blocks that can span multiple layers. The challenge for physical design is to automate the process of placing blocks in multiple layers.

To address these challenges, we make the following contributions:

- 3D Architectural Blocks: We propose *port partitioning*, an approach to place architectural blocks like register files, issue queues, and caches in multiple silicon layers. We compare port partitioning with wordline/bitline partitioning [30] with respect to area, timing, power, and required vertical interconnect.

- 3D Microarchitectural and Physical Design Co-Optimization: We extend the MEVA framework [6] to handle fine-grain 3D exploration. Our modified framework can automatically choose between 2D and 3D implementations of a given block. Given a frequency target, an architectural netlist, and a pool of alternative block implementations, this framework can find the best solution in terms of performance (in BIPS), temperature, or both.

- 3D Design Driver Exploration: Using our modified framework, we explore the design space of different partitioning schemes for a particular design driver architecture, using one to four layers of silicon. In addition to exploring the use of single layer and multilayer blocks, we consider growing the sizes of different architectural structures, using the timing slack from vertical integration. In some cases, the timing slack can enable the use of larger instruction or scheduling windows, or larger caches.

In addition to helping latency, this reduction in wire RC delay can reduce power dissipation. However, the stacking of components can adversely impact the temperature of the microprocessor. It is therefore essential for any study using vertical integration to make use of accurate temperature modeling to demonstrate the effectiveness of any architecture. All of our explorations are enhanced with a state-of-the-art, accurate, temperature simulator tool. We also consider automated thermal via insertion to help mitigate the impact of temperature.

The rest of the paper is organized as follows: We review the prior work on 3D integration technology, microarchitectural exploration techniques, and block modeling in Section 2. Next, we detail and evaluate our 3D architectural blocks in Section 3. Our 3D block placement enhancements are detailed in Section 4. We finally explore a design driver microarchitecture in Section 5 and then conclude in Section 6.

# 2 Related Work

In this section we focus on the most relevant prior work to our study.

## 2.1 3D Technologies

While a number of 3D IC fabrication technologies have been proposed [17, 22, 19], we consider the use of wafer bonding [1, 2, 7] in this study. In this technology, fully processed wafers are bonded together, and devices are fabricated on these wafers. Interlayer vias that connect different layers are etched after metalization and prior to wafer bonding. Two main kinds of wafer bonding strategies have been evaluated in

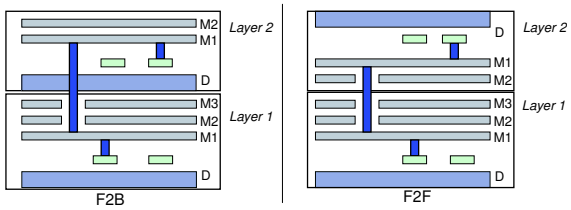



*Figure 1: Face-to-Back and Face-to-Face integration technologies*

prior work [2, 7]: Face-to-Back (F2B) placement and Face-to-Face (F2F) placement (Figure 1). Vias in F2B cut through device layers in addition to metal layers. In F2F placement, the top device layer is flipped to face the lower device layer. Metal layers are placed between the facing device layers. Hence, vias cut through metal layers only. However, F2F cannot scale beyond two layers without also employing F2B layers.

## 2.2 3D Microarchitectural Exploration

MEVA-3D [6] is an automated exploration framework that can explore a 3D design space for an optimal placement of 2D architectural blocks into multiple device layers. MEVA-3D optimizes a cost function that is configured to weigh latencies of critical microarchitectural loops, temperature, and die area. The critical loop latency is the sum of individual block latencies along the loop and inter-block wire latencies. Critical loop latencies relate to performance (IPC) as in [28]. The algorithm returns a floorplan with the best overall performance, temperature and die area for a given target frequency. MEVA-3D leverages SimpleScalar [4] to validate its performance estimate. MEVA-3D can also perform automated thermal via insertion to help mitigate areas of high power density. However, MEVA-3D does not currently support the exploration of 3D designs using 3D blocks.

Ozturk et. al. proposed a 3D topology optimization algorithm [21]. The algorithm considers the optimal placement of a few processor cores that are associated with a large number of storage blocks. The algorithm is able to improve performance by placing these cores and blocks in 3D so that the cores are closer to their most frequently accessed storage blocks. However, this algorithm does not consider the placement of actual microarchitectural blocks such as the ALU, issue queue, branch predictor, etc, and does not consider the latency reduction of critical microarchitectural loops. The algorithm is also not able to explore the placement of 3D-designed blocks.

## 2.3 2D and 3D Block Modeling

Prior work has provided block models for various architectural structures including caches [31], register files [9, 23], and wakeup and select logic [23]. CACTI [31, 25, 26] is an analytical model that provides timing, area, and power results for different cache configurations. CACTI models different levels of associativity, multiporting, sub-banking, and ideally scales to different feature sizes using 0.80$\mu$m cache data. Tsai et al [30] extended CACTI to explore 3D cache designs. However, they only consider folding blocks by wordlines or bitlines, and not by port partitioning. In addition, they do not explore the impact of this 3D design on the overall microarchitecture (i.e. performance, temperature, layout), or the impact of 3D stacking on area in general. Puttaswamyy et al. [14] showed the delay benefit and the reduction of power consumption in a stacked cache design by bank-stacking or array-splitting. There has been no prior work that explores partitioning cache ports. Palacharla et al [23] built detailed transistor-level models for critical structures in dynamically scheduled processors, analyzing critical timing paths and the scalability of these structures. However, this study is limited to single layer structures.

# 3 3D Architectural Block Design and Modeling

To reduce intra-block interconnect latency, we evaluate two main strategies for designing blocks in multiple silicon layers: *block folding* and *port partitioning*. Block folding implies either a vertical or horizontal folding of the block - potentially shortening the wirelength in one direction. Port partitioning places the access ports of a structure in different layers - the intuition here is that the additional hardware needed for replicated access to a single block entry (i.e. a multiported cache) can be distributed in different layers, which can greatly reduce the length of interconnect within each layer. In this section, we describe the use of these strategies for the issue queues and various cache-like blocks in our design driver architecture.

## 3.1 Issue Queues

The issue queue is a critical component of out-of-order microprocessor performance and power consumption. Recent research [6] has shown that every additional pipeline stage of latency seen in the scheduling loop causes an average 5% performance degradation. Moreover, Folegnani and Gonzalez [10] have found that the issue queue is responsible for an average 25% of a processor's total power consumption.

The issue queue stores renamed instructions and performs out-of-order instruction scheduling. The issue queue we studied in this paper is based on Palacharla's implementation [23]. There are two main stages of issue queue functionality: the wakeup stage where tags from completing register values are compared against input register tags stored in issue queue entries, and a selection stage where ready instructions (as determined by the wakeup stage) are selected for execution.

Each issue queue entry must track and compare the input register tags required by a given instruction in that entry. Figure 2 shows a single CAM cell used to store one bit of a register tag for an issue queue entry. Assuming that at most four register values can be written back each cycle, and at most four new instructions can enter the issue queue each cycle, an individual cell would have four different 1-bit tags to compare against and have four write ports. In a processor with a 128-entry physical register file, register tags are 7-bits. Therefore each row would need seven CAM cells for each operand, for a total of fourteen CAM cells. In general, an n-entry issue queue has n such rows.

In the wakeup stage, the match lines for each issue queue entry are precharged high and the tag lines are driven with the register tags of completed instructions. A match line only remains high if the register tag stored at the issue queue entry is the same as a certain one of the register tags driven on the tag lines. If any match line for a given input register remains high, the ready bit for that operand is set in the issue queue. Once both ready bits are set, the operand is eligible for issue (i.e. has woken up). In this stage, most of the delay comes from tag broadcasting and matching.

In the selection stage, the select logic picks instructions to execute [23] among all instructions that are eligible for issue.

For example, a selection tree for a 32-entry issue queue consists of three levels of arbiters. Each arbiter takes four input requests (i.e. four eligible instructions) and grants one request (i.e. selects one eligible instruction). In general, an N-entry issue queue needs a selection tree of level $L = log_4 N$.

In the issue queue, the delay due to wakeup logic contributes a large portion of the overall delay. Our simulations show that wakeup takes about 60% of the delay in a 32-entry issue queue with four incoming register tags to compare against, and four access ports. A significant contributor to delay is the wire latency of the tag bits and match lines. A 3D integrated issue queue can significantly reduce the length of these wires.

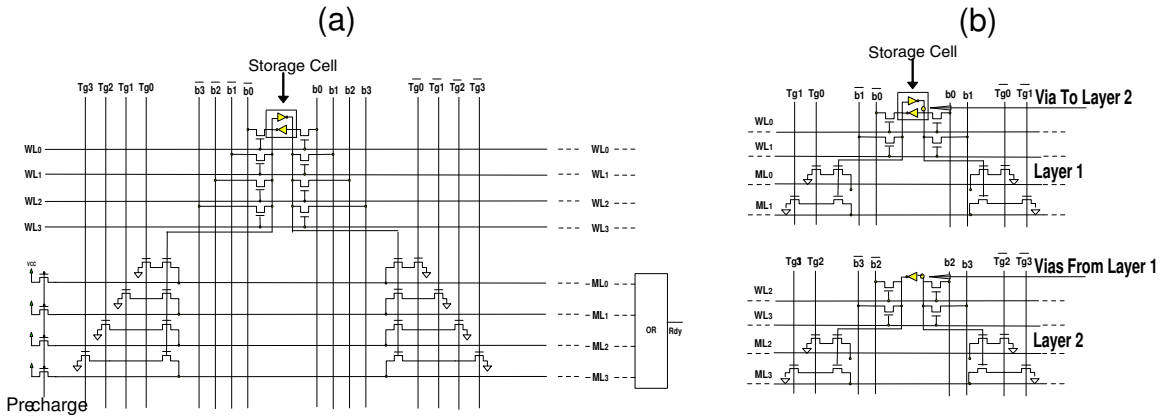



*Figure 2: (a): A Single IQ Cell with Four Tag Lines and Four Access Ports. Over 99% of the area is occupied by tags and access ports.*

*(b): Port Partitioning. Tags and access ports are distributed into two layers. Width and height of each bit are reduced by half, and area by 75%.*

### 3.1.1 3D IQ Design: Block Folding

One way to reduce tag line wire delay is to fold the issue queue entries and place them on different layers. Figure 3 (a) shows a single layer issue queue with four incoming register tags that are compared against entries in the issue queue. In Figure 3(b), the issue queue is folded into two sets and they are stacked in two layers. This approach effectively shortens the tag lines.

### 3.1.2 3D IQ Design: Port Partitioning

In an issue queue with four tag comparison ports and four read/write ports, as shown in Figure 2(a), most of the silicon area is allocated to ports. The wire pitch is typically five times the feature size [25, 26, 23]. For each extra port, the wire length in both X and Y directions is increased by twice the wire pitch [25, 26]. On the other hand, the storage, which consists of 4 transistors, is twice the wire pitch in height, and has a width equal to the wire pitch. Hence, in a cell as shown in Figure 2(a), the storage area is less than 1% of the total area, while tags and access ports occupy over 99% of the total area.

One strategy to attack the tag and port requirements is port partitioning, which places tag lines and ports on multiple layers, thus reducing both the height and width of the issue queue. The reduction in tag and matchline wire length can help reduce both power and delay. The selection logic also benefits from this, as the distance from the furthest issue queue entry to the arbiter is reduced. This will speed up the comparison and also reduce power consumption.

## 3.2 Caches

Caches are commonly found architectural blocks with regular structures - they are composed of a number of tag and data arrays. Figure 4(a) demonstrates a high level view of a number of cache tag and data arrays connected via address and data buses. Each vertical and horizontal line represents a 32-bit bus – we assume two ports on this cache, and therefore the lines are paired. Each box of the figure is a tag or data array, which is composed of a mesh of horizontal wordlines and vertical bitlines. Every port must have a wordline for each cache set and a pair bitlines for each bit in a cache set. The regularity of caches means that their components can easily be subdivided – the tag and data arrays for example can easily be broken down into

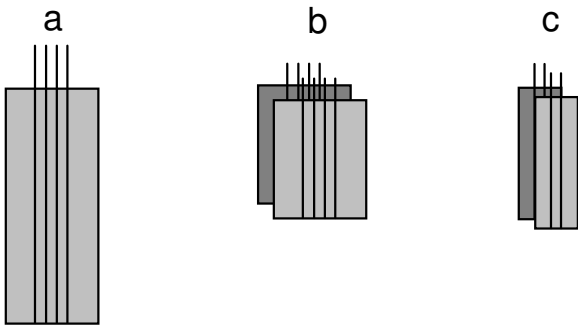



*Figure 3: Issue Queue Partitioning Alternatives: (a) An issue queue with 4 tag lines. (b) Block Folding: dividing the issue queue entries into two sets and stacking them. The tags are duplicated in every layer. Only the X-direction length is reduced. (c) Port Partitioning: the four tags are divided into two tags on each layer. Both X and Y direction lengths are reduced.*

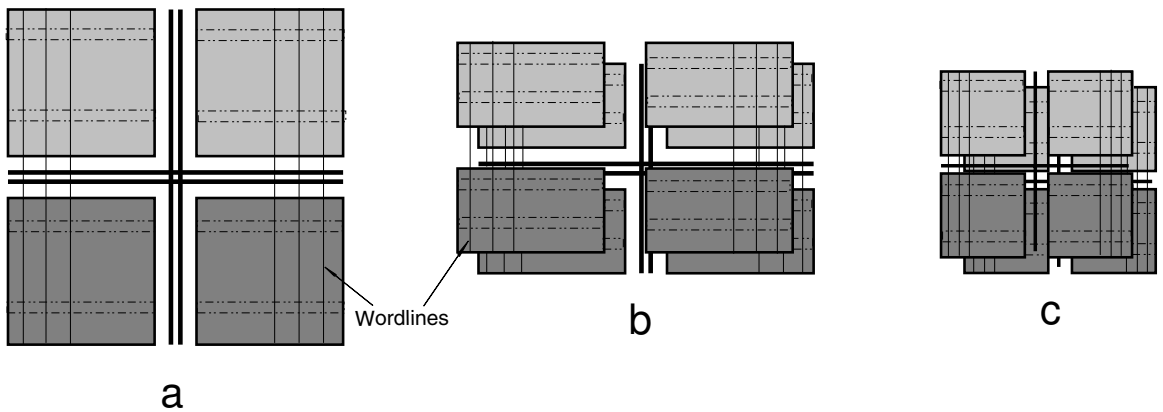



*Figure 4: Cache Block Alternatives (a) A 2-Ported Cache: the two lines denote the input/output wires of two ports. (b) Wordline Folding: Only Y-direction length is reduced. Input/output of the ports are duplicated. (c) Port Partitioning: Ports are placed in two layers. Both X and Y direction length are reduced.*

subarrays. We make use of CACTI [26] to explore the design space of different subdivisions and find an optimal point for performance, power, and area.

### 3.2.1 3D Cache Design: Block Folding

Prior research [30] looks into two folding options: wordline and bitline folding. In the former, the wordlines in a cache sub-array are divided and placed onto different silicon layers. The wordline driver is also duplicated. The gain from wordline folding comes from the shortened routing distance from predecoder to decoder and from output drivers to the edge of the cache.

Similarly, bitline folding places bitlines into different layers. This approach needs to duplicate the pass transistor. The sense amplifier can be duplicated to improve timing performance at a cost of increased power consumption. The cost is significant because sense amplifiers can make up a significant portion of total cache energy consumption. The other approach is to share sense amplifiers across layers, but this dramatically reduces the improvement in timing.

Our investigation shows that wordline folding has a better access time and lower power dissipation in most cases compared with a realistic implementation using bitline folding. In this paper, we only present results using wordline folding.

### 3.2.2 3D Cache Design: Port Partitioning

The port partitioning strategy that we proposed for the issue queue can also be leveraged for caches. For example, a 3-ported structure would have a port area to cell area ratio of approximately 18:1. Hence, there is a significant advantage to partitioning the ports and placing them onto different layers. In a two layer design, we can place two ports on one layer, one port and the SRAM cells on the other layers. The width and height are both approximately reduced by a factor of two, and the area by a factor of four.

## 3.3 Other Cache-Like Architectural Blocks

Register files are similar to caches, sharing the regularity of a cache. We therefore adapt our CACTI to model this structure as well. However, they are not associative and typically have more ports than caches do. Register files dissipate relatively large amounts of power due to their porting requirements, and the size of the physical register file can constrain the size of the instruction window in a dynamically scheduled superscalar processor. We will consider the same folding schemes for the register files as we used for caches.

The register mapping units, load-store queue, and branch predictors can be approximated using only the data array portion of the cache.

## 3.4 Modeling Methodology

We assume a supply voltage of $1.0V$ and a $70nm$ process technology. Transistor and wire scaling parameters are derived from [30, 18], and we assume copper interconnect in our simulation. Further transistor parameters are obtained from [5]. The 3D via resistance is estimated to be $10^{-8}\Omega cm^2$ [30]. The height of the 3D vias is assumed to be $10\mu m$ per device layer. Current dimensions of 3D via sizes vary from $1\mu m \times 1\mu m$ to $10\mu m \times 10\mu m$ [30, 8]. As 3D technology advances, the 3D via size will decrease even further. In this study, we assume the via pitch is $1.4\mu m$. An area of $0.7\mu m \times 0.7\mu m$ is reserved for each 3D via for the upper layers in F2B technology.

We have modified 3D-CACTI [30] to model caches and cache-like structures. First, we add port partitioning to 3D-CACTI in addition to wordline/bitline folding. Second, we add area estimation, including the area impact of 3D vias on the transistor layer. Both 3D bonding technologies are available: F2B and F2F. We validated our modifications to 3D-CACTI with HSpice.

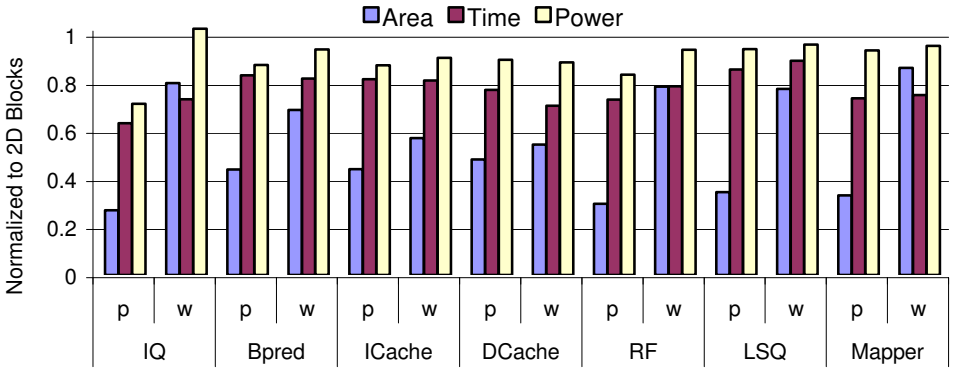



*Figure 5: The improvement in area, power and timing for dual layer vertical integration.*

We implemented our issue queue models using HSpice to obtain accurate timing and power data. The area of the issue queue is approximated by 3D-CACTI using a similarly sized cache. Our 2D issue queue is derived from Palacharla et al's model [23].

### 3.5 3D Block Performance

Figure 5 demonstrates the effectiveness of 3D block design on area, power, and timing for dual layer F2F blocks. The y-axis is normalized to the area of a single layer baseline block. The x-axis represents different folding techniques for each architectural block investigated. The letters in the label of a bar represents the type of folding: either port partitioning (PP) or block folding (BF). All results are shown normalized to the 2D implementation of the block. In F2F technology, the via starts from the surface of one layer and ends on the surface of the other layer. Therefore, vias do not impact the layout of transistors.

For the issue queue (IQ), delay is reduced by 27% with BF. PP sees even more improvement (37% reduction in delay). PP reduces both tag wire lengths and match wire lengths, and wire lengths to the selection logic. On the other hand, BF only reduces tag wire lengths. The match wire lengths are even increased due to 3D via insertions for every tag and bit line. As a result, we observe over 70% reduction in area for PP, with only a 20% reduction for BF. Note that the area shown is the maximal area in any one layer for that block, and while the footprint of the block may be reduced, the sum of the area occupied in all layers may actually increase relative to the 2D baseline.

The power consumed in CMOS circuits is represented as $P = 0.5 * a * f * C * Vdd^2$, where $f$ is the clock frequency, $a$ is the activity factor, $Vdd$ is the supply voltage and $C$ is the switching capacitance. The power consumption rate is proportional to the switching capacitance. In BF, although tag wire lengths for each layer are reduced, the tag wires are duplicated on different layers. The aggregate wire length is still the same. In addition, there is an increase in match line lengths mentioned above. Thus, the total switching capacitance is slightly increased due to the increased total wire length. As a result of this, the power consumption of BF is slightly increased. On the other hand, PP is able to reduce power consumption by 29%.

For the caches and cache-like structures, PP is extremely effective in heavily ported structures. For example, the register file with PP sees a 27% reduction in delay, a 17% reduction in power, and an impressive 70% reduction in area. However, for structures with fewer ports, BF can be more effective. The data cache sees a 30% reduction in delay with BF, and a 23% reduction in delay with PP. While PP does reduce both

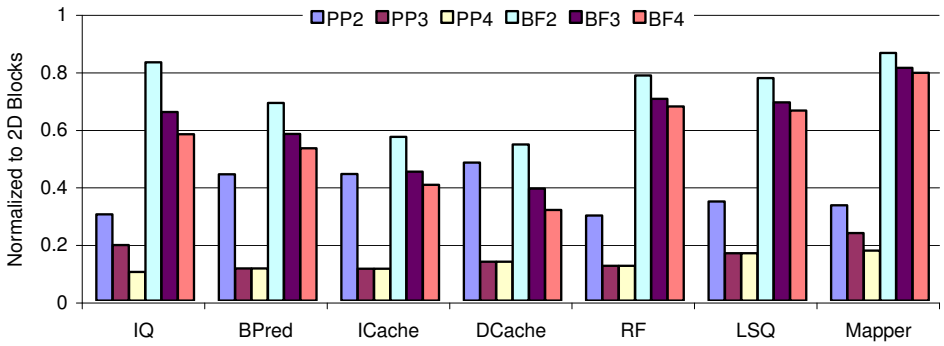



*Figure 6: The improvement in area for multilayer F2B vertical integration.*

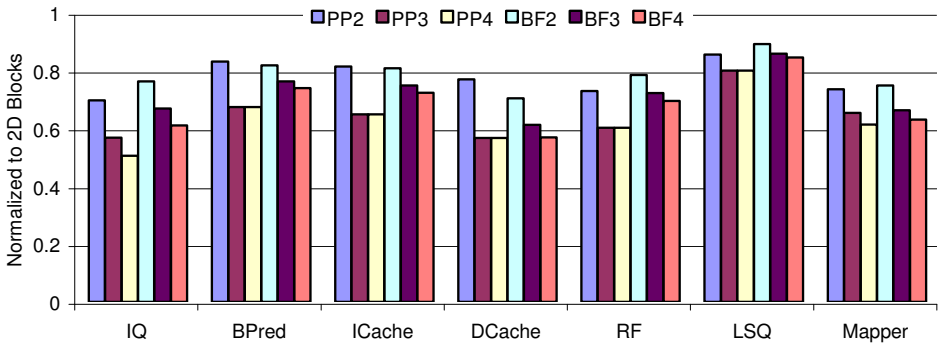



*Figure 7: The improvement in timing for multilayer F2B vertical integration.*

wordline and bitline length, this reduction is proportional to the number of ports that can be partitioned to other silicon layers. For structures with very few ports, BF is able to reduce wordline length more than PP. Hence in structures that have significant wordline delay, the overall reduction in delay with BF can be greater than PP.

The diversity in benefit from these two approaches demonstrates the need for a tool to flexibly choose the appropriate implementation based on the constraints of an individual floorplan.

### 3.5.1 Scaling to Multiple Silicon Layers

For a dual layer implementation, F2F is able to outperform F2B since the 3D vias in F2B impact the silicon footprint in the top silicon layers. For example in the PP results, the F2B area is about 5% larger than that of F2F due to the increased silicon footprint. The delay and power consumption are larger than those of F2F as well. However, F2B allows more layers to be stacked. It may be possible to stack two F2Fs in back to back fashion; however, we do not consider this alternative in this paper.

Figures 7, 8, and 6 show timing, power, and area results (respectively) with F2B blocks for two, three, and four layers of silicon. All measurements are normalized to the performance of a single layer block. In general, we observe that the reduction of area, power and delay is further increased as the number of layers is increased.

For the issue queue (IQ) with PP, area reduction increases to 80% with 3 layers, and to 90% with 4

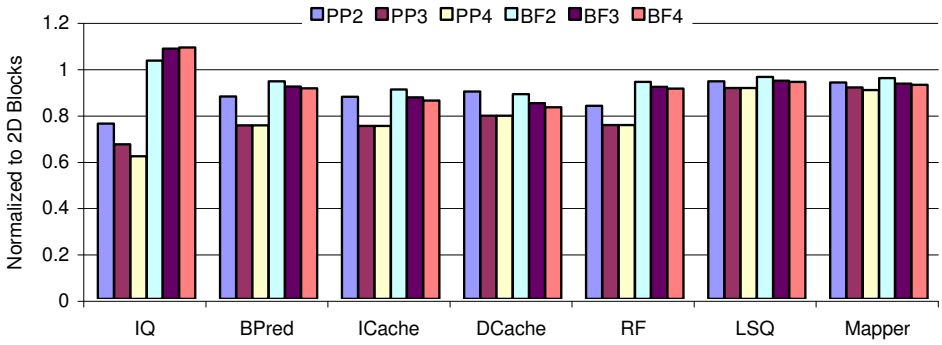



*Figure 8: The improvement in power for multilayer F2B vertical integration.*

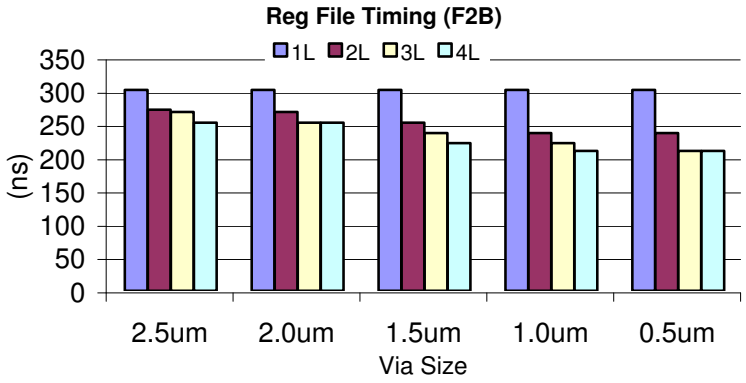



*Figure 9: Impact of Via Size on Timing using F2B, Port Partitioning*

layers. Reduction in issue queue delay increases to 43% with 3 layers, and to 50% with 4 layers. Reduction in power consumption grows as high as 38% with 4 layers.

For the issue queue with block folding, there is less reduction in area and delay with additional layers. However, the impact on match line wire length from stacking more layers increases the power consumption for folding to 9% with 4 layers.

### 3.5.2 Impact of 3D Bonding Technology

3D via size has rapidly scaled down as 3D bonding technology has advanced. 3D via size has reduced from $10\mu m$ to $1.75\mu m$ in MIT Lincoln Laboratory's 3D process technology [15] at $180nm$. We expect the 3D via size to continue to scale at smaller feature sizes. In this paper, we have assumed a $0.7\mu m$ via size for a $70nm$ feature size.

To demonstrate the impact of scaling via size, we plot the performance of the register file for via sizes ranging from $2.5\mu m$ to $0.5\mu m$ in $70nm$ technology. The via pitch is twice the via size. The register file has four read ports and four write ports. A single cell size is approximately $5.6\mu m$ x $5.6\mu m$. In F2B bonding technology, 3D vias occupy silicon area in all layers except for the bottom layer. Taking 2-Layer partitioning as an example, when via size is $2.5\mu m$, the best solution is to place seven ports in the bottom layer, and one port in the top layer, which only slightly reduces the wirelength. When the via size is scaled to $0.5\mu m$, the best solution places four ports in each layer. The wirelength is almost cut in half in both X and Y directions.

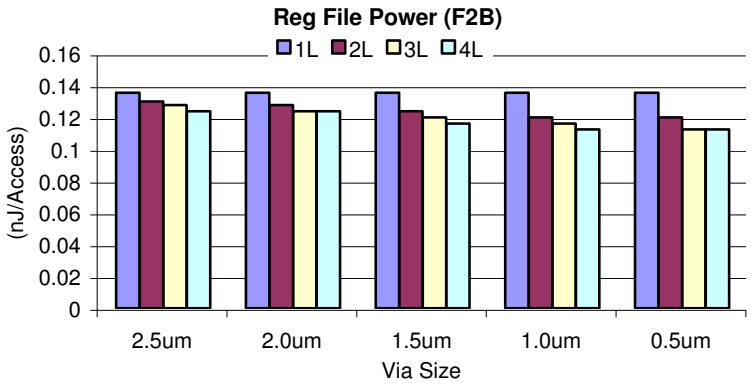



*Figure 10: Impact of Via Size in Power using F2B, Port Partitioning*

As shown in Figures 9 and 10, the larger reduction in wirelength reduces both delay and power as the via size is scaled from $2.5\mu m$ to $0.5\mu m$.

# 4 3D Block Placement

Microprocessor throughput, as measured in IPC, is influenced by the latency of critical architectural loops such as the scheduling loop, branch resolution loop, inter-cluster communication loop, etc [28]. Vertical integration can help to reduce the latency of these critical loops. Critical loops differ in the magnitude of their impact on throughput, and therefore the exploration of the use of vertical integration on microprocessor design requires consideration for both physical design and architecture. Existing work on this type of co-design exploration [6] has only explored the use of vertical integration to reduce inter-block latency in these critical loops. However, as demonstrated in section 3, there is tremendous potential for vertical integration to reduce the latency of blocks along critical loops. In this section, we detail our modifications to the co-design framework of [6].

## 4.1 MEVA-3D Flow

MEVA-3D [6] is an automated physical design and architecture performance estimation flow for 3D architectural evaluation which includes 3D floorplanning, routing, interconnect pipelining, automated thermal via insertion, and associated die size, performance, and thermal modeling capabilities.

First, MEVA-3D takes a microarchitectural configuration, a target frequency, architectural critical path sensitivities, and power density estimates and uses 2D/3D floorplanning to optimize for performance and temperature. Then routing and thermal via planning are performed to provide physical design information to our microprocessor simulation. Critical loop latencies are passed from the floorplanner to the simulator for accurate determination of performance. MEVA-3D makes use of the SimpleScalar [4] simulator to obtain performance in IPC and utilization counts of individual blocks.

## 4.2 Enhancements to MEVA-3D

MEVA-3D currently only considers 2D architectural blocks. We make the following modifications to extend it to 3D blocks. In the following section, we will make use of this modified framework to explore an architectural design driver.

| | |
|---|---|
| Processor Width | 6-way out-of-order superscalar, two integer execution clusters |
| Register Files | 128 entry integer (two replicated files), 128 entry FP |
| Data Cache | 8KB 4-way set associative, 64B blocksize |
| Instruction Cache | 8KB 2-way set associative, 32B blocksize |
| L2 Cache | 4 banks, each 128KB 8-way set associative, 128B blocksize |
| Branch Predictor | 8K entry gshare and a 1K entry, 4-way BTB |
| Functional Units | 2 IntALU + 1 Int MULT/DIV in each of two clusters<br>1 FPALU and 1 MULT/DIV |

*Table 1: Architectural parameters for the design driver used in this study.*

#### 4.2.1 Architectural Alternative Selection

3D component design gives us different configurations for each component: the number of layers that the component will occupy. When we choose another configuration for a component, the dimensions, timing characteristics, and power values change as well, which usually results in a significant change to the floorplan. In order to explore the combinations of the different configurations, we introduce a new type of move in the optimization approach, called architecture alternative selection. When a new configuration is selected, the 2D dimensions, layer information, and delay information are updated. Accordingly, the distribution of the power, including leakage, is updated for all blocks in the design. The new packing result may be accepted or rejected depending on the cost function evaluation.

#### 4.2.2 Cube Packing Engine

Because of the limitation of the packing engine used in MEVA-3D, each component can only occupy one layer, and therefore 3D components are not allowed in the original MEVA-3D flow. To enable the packing of 3D components which may occupy more than one layer, we constructed a new packing engine which is a true 3D packing engine – 3D components in our design can be treated as cubic blocks to be packed in 3D space. The dimension of the block in the Z direction represents the layer information. The 3D packing algorithm is extended from the CBL floorplanner [16].

Our 3D CBL(3-Dimensional Corner Block List) packing system represents the topological relationship between cubic blocks with a triple (S,L,T), where each element is a list. List S records the packing sequence by block names. List L and T represent the topological relationship between cubic blocks in terms of covering other packed blocks.

We use simulated annealing to optimize the cubic packing. The number of layers is given as a constraint on the maximal height in the Z direction of the packing. We extended the floorplanner to optimize chip area, performance (using microarchitectural loop sensitivities), and temperature at the same time.

## 5 Microarchitectural Exploration

In this section, we use the modified MEVA framework to investigate the ability of vertical integration to reduce both intra-block and inter-block architectural latencies.

We constructed a design driver based loosely on the Alpha 21264 [13], and along with the architectural blocks from Section 3 (functional unit blocks are based on [6]), we feed this driver into our modified version of MEVA-3D. The architectural parameters are shown in Table 1. We measure architectural performance on all 26 programs of the SPEC CPU2000 suite.

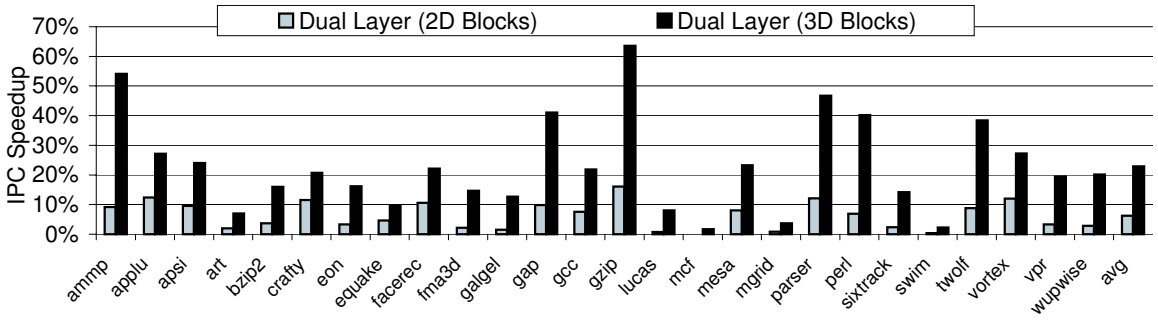



*Figure 11: Performance speedup for dual silicon layer architectures relative to a single layer architecture.*

Figure 11 presents performance results relative to a single layer design driver. The first bar represents the benefit from using two layers of silicon with 2D blocks (as in [6]) and the second bar represents the benefit from using two layers of silicon with 3D blocks. All three configurations (single layer, dual layer 2D blocks, dual layer 3D blocks) are running at 4GHz. On average, the use of 2D blocks in a two layer design improves performance by 6%. Since the blocks themselves do not take advantage of vertical integration, any performance gain can only come from a reduction in the inter-block wire latency. For example, the branch misprediction loop has a total latency of 815ps at 4GHz for a single layer design – 238ps of this total latency is from inter-block wire delay. When using 2D blocks in two layers, this inter-block wire delay is reduced to only 63ps. However, the overall reduction in path delay is not enough to reduce the loop by a cycle of our 4GHz clock. Thus, while timing slack is certainly increased, the benefit of this has not been exploited in Figure 11. When we allow MEVA-3D to select 3D block alternatives, we see a performance improvement of 23% on average over the single layer architecture. This can be attributed to the ability of 3D blocks to reduce the intra-block latency of critical processor loops.

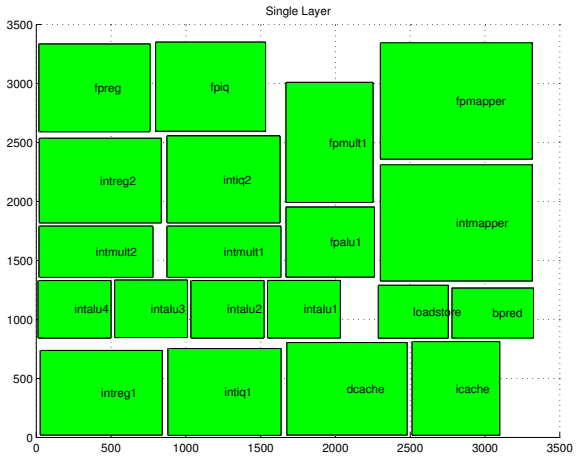
We show floorplans for all three architectures in Figure 12. The single layer design occupies $3.4 \times 3.4mm^2$ in one silicon layer. The dual layer design with 2D blocks occupies $2.8 \times 2.8mm^2$ in each silicon layer. The dual layer design with 3D blocks occupies $2.3 \times 2.3\ mm^2$ in each silicon layer.

Temperature issues are considered to be a major concern for vertical integration. Therefore, an accurate and fast thermal simulation framework was very crucial for our experimental analysis. We used the finite element method (FEM) based CFD-ACE+ temperature simulator [24]. Further details on heat sink and thermal parameters we used can be found in [24]. Figure 13 presents the average core temperature for the single layer architecture (shown at left) and the dual layer architecture with 3D blocks (the hottest layer is shown at right). The average and maximum temperature for the single layer architecture was $30.6°$C and $32.7°$C. The average and maximum temperature for the dual layer architecture with 2D blocks was $30.6°$C and $32.6°$C. The average and maximum temperature for the dual layer architecture with 3D blocks was $30.3°$C and $34.1°$C.

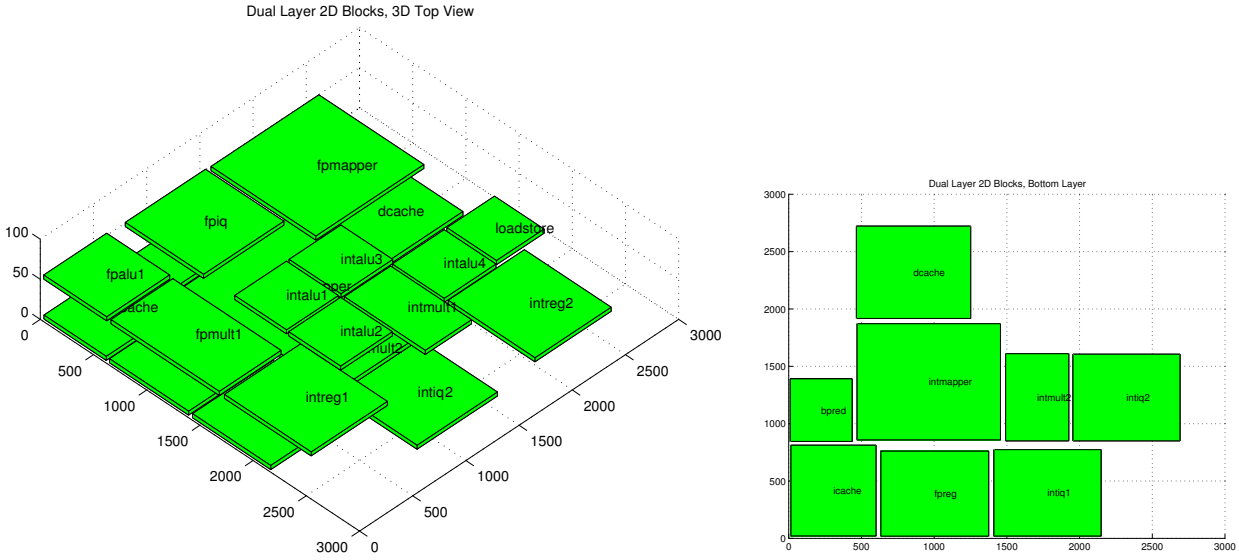
Thermal vias can help to relieve thermal problems in 3D microarchitectures. We used the algorithm proposed in [11] for thermal via insertion. In our multi-layer designs, we designate 5% of the area as dead space on each layer, which provides sufficient space for thermal vias.

## 5.1 Scaling Architectural Sizes

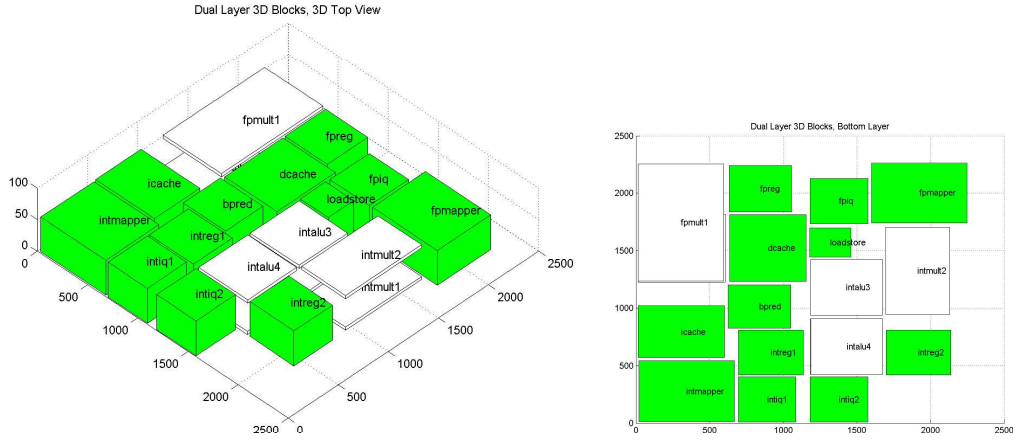
Even in the 3D block architecture, there are still cases where we are able to increase the timing slack within a given cycle of a critical loop, without actually reducing the number of cycles in that critical loop. Figure 14 presents one approach to leveraging this extra slack: we double the size of the data cache, issue queue, and register file.




(a) A single Layer Architecture




(b) Dual Layer Architecture with 2D-only Blocks




(c) Dual Layer Architecture with 2D and 3D Blocks. Dark blocks occupy two layers and white blocks occupy one layer.

*Figure 12: Floorplans for the best architectural configuration as determined by our modified version of MEVA-3D*

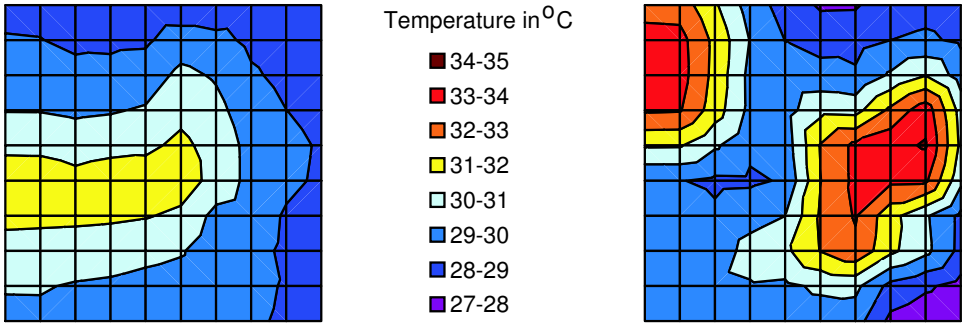



*Figure 13: Average core temperature for the single layer architecture (shown at left) and the dual layer architecture with 3D blocks (the hottest layer is shown at right).*

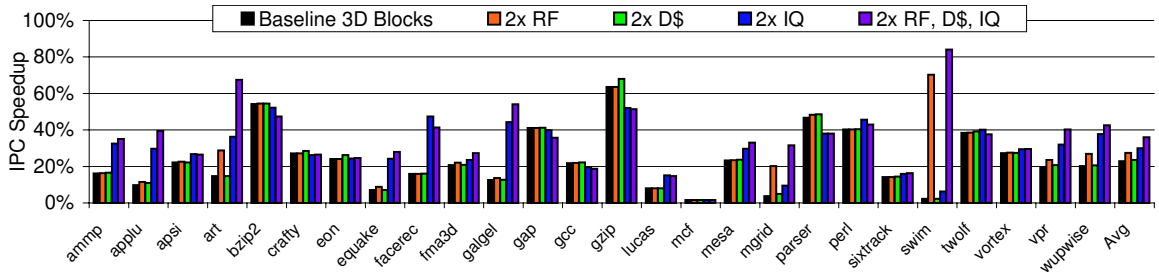



*Figure 14: Performance when doubling critical resources.*

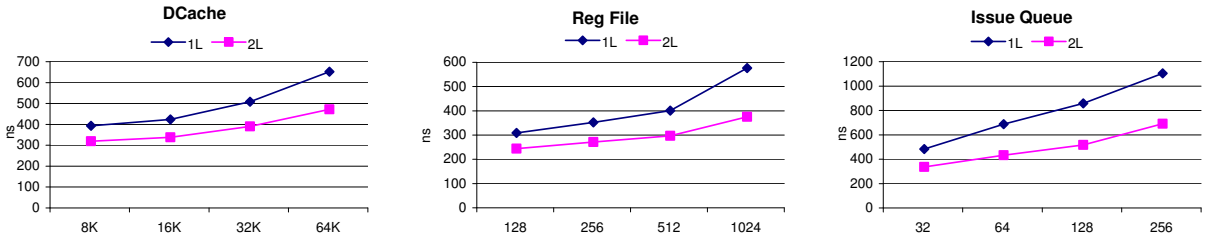



*Figure 15: Latency Impact of Vertical Integration when Scaling the Size of Three Critical Blocks*

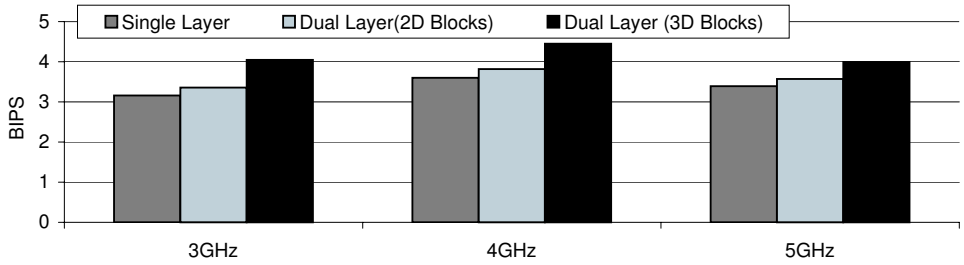



*Figure 16: BIPS performance for different clock frequencies.*

Figure 15 shows the timing performance when three structures are scaled from the default size to 16 times larger. As shown in the figure, using 3D integration technology, the access latency of double-sized structures is still less than in 2D. The register file and data cache can even quadruple their sizes while still outperforming the default blocks in 2D. In this paper, we limit our study to doubled sizes.

As shown in Figure 14, the performance is increased by an additional 5% with a doubled cache, an additional 1% with a doubled register file, and an additional 7% with a doubled IQ. The best performance is observed when doubling the size of all three structures. Overall, there is a 36% gain over the 2D architecture and a 13% gain over the 3D architecture with our default block sizes.

These larger structures will dissipate more power than regular-sized 3D blocks. But despite the increase in power, the increased area of these larger designs saw an average slight decrease in temperature of $0.8°$C for the case where all three resources were doubled. The maximal temperature in this case was $34.1°$C.

## 5.2 Frequencies

The results presented so far feature a 4GHz clock frequency. Figure 16 demonstrates the performance in BIPS when using different clock frequencies: 3GHz, 4GHz, and 5GHz. The first bar is the single layer architecture, the second bar is the dual layer architecture with 2D blocks, and the final bar is the dual layer architecture with 3D blocks.

At 3GHz, the larger latency of wire using 2D architecture or 2D blocks is better tolerated by the more forgiving clock rate – but the use of 3D blocks can still provide a 10% gain over a single layer architecture. The overall performance at 5GHz decreases slightly from 4GHz due to the lengthened critical loops at a higher frequency.

## 5.3 Number of Layers

In this subsection, we demonstrate the performance of vertical integration when scaling beyond two silicon layers. Figure 17 illustrates this gain for 4GHz architectures. Due to the challenge in scaling to more layers with F2F blocks, we only use F2B in this study.

The performance of two layers in this figure is slightly worse than in Figure 11. The F2B bonded blocks used in this section perform slightly worse than F2F blocks because of the impact of 3D vias on the silicon layer.

The first bar shows the performance when using only 2D blocks, and as shown in the figure, there is little gain from scaling the number of silicon layers for this design driver. Without tackling intra-block latency, even the near elimination of inter-block latency can only improve performance by so much. However, the gain from 3D blocks over the single layer architecture grows from 22% at two layers to 28% at four layers. Although inter-block wire latencies have been nearly eliminated in two layers, the latency reduction in four-layer microarchitectural blocks further reduced cycles in critical loops.

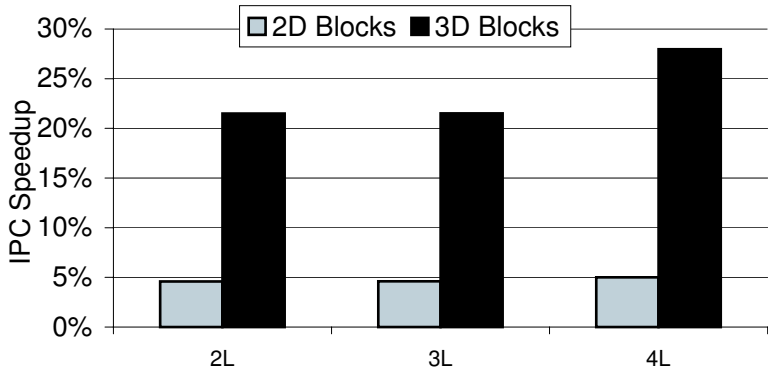



*Figure 17: Performance when scaling the number of silicon layers.*

There is little gain from two layers to three layers using 3D blocks. As shown in Figure 7, many architectural blocks have little or no reduction in latency from two layers to three layers. Furthermore, while there was an increase in slack for many critical loops, there was no overall reduction in cycles for these critical loops. However, it would certainly be possible to leverage this slack in other ways.

# 6 Summary

Vertical integration has tremendous potential to reduce both inter-block and intra-block wire latency. We have proposed and evaluated tag partitioning for the issue queue, for caches, and for cache-like blocks. And we have enhanced the MEVA-3D exploration framework to evaluate the use of 3D blocks in multiple layers of silicon. When using two layers of silicon with 3D blocks, we see an average 36% improvement in performance over a single layer architecture and 29% improvement in performance over two layers with single layer blocks, for the architectural design driver we explored. Temperature is kept below $40^{\circ}\mathrm{C}$ using a two heat sink F2F design.

# References

[1] K. Banerjee, S. Souri, P. Kapur, and K. Saraswat. 3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration. In *In Proc. of the IEEE, 89(5):602-633*, May 2001.

[2] B. Black, D. W. Nelson, C. Webb, and N. Samra. 3d processing technology and its impact on ia32 microprocessors. In *Proc. Of ICCD, pp.316-318*, 2004.

[3] E. Borch, E. Tune, S. Manne, and J. Emer. Loose loops sink chips. In *Proceedings of the Eighth International Symposium on High-Performance Computer Architecture*, 2002.

[4] D. C. Burger and T. M. Austin. The simplescalar tool set, version 2.0. Technical Report CS-TR-97-1342, U. of Wisconsin, Madison, June 1997.

[5] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu. New paradigm of predictive mosfet and interconnect modeling for early circuit design. In *Proc. of Custom Integrated Circuit Conference*, 2000.

[6] J. Cong, A. Jagannathan, Y. Ma, G. Reinman, J. Wei, and Y. Zhang. An automated design flow for 3d microarchitecture evaluation. In *Proc. Asia and South Pacific Design Automation Conf*, January 2006.

[7] S. Das, A. Chandrakasan, and R. Reif. Design tools for 3-D integrated circuits. In *Proc. Asia and South Pacific Design Automation Conf., pp. 53-56*, January 2003.

[8] S. Das, A. Fan, K. Chen, and C. Tan. Technology, performance, and computer-aided design of three-dimentional integrated circuits. In *Proc. International Symposium on Physical Design*, April 2004.

[9] K. Farkas, N. Jouppi, and P. Chow. Register file design considerations in dynamically scheduled processors. In *Proceedings of the Second International Symposium on High Performance Computer Architecture*, 1995.

[10] Daniele Folegnani and Antonio Gonzalez. Energy-effective issue logic. In *Proceedings of the 28th Annual International Symposium on Computer Architecture (ISCA'01)*, pages 230–239. ACM Press, 2001.

[11] J.Cong and Y.Zang. Thermal-driven multilevel routing for 3-D ICs. In *Asia Pacific Design Automation Conference*, pages 121–126, 2005.

[12] K.Banerjee, S.Souri, P.Kapur, and K.C. Saraswat. 3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration. In *IEEE Special Issue Interconnections - Addressing The Next Challenge of IC Technology, Vol. 89, No. 5*, pages 602–633, 2001.

[13] R.E. Kessler, E.J. McLellan, and D.A. Webb. The Alpha 21264 microprosessor architecture. In *International Conference on Computer Design*, December 1998.

[14] K.Puttaswamy and G.H.Loh. Implementing caches in a 3d technology for high performance processors. In *International Conference on Computer Design*, 2005.

[15] MIT Lincoln Laboratory. Mitll low-power fdsoi cmos process: Design guide. March 2006.

[16] Y. Ma, X. Hong, and C.K. Cheng S. Dong. 3D CBL: An efficient algorithm for general 3-dimensional packing problems. In *In IEEE International Midwest Symposium on Circuits and Systems*, August 2005.

[17] M.A.Crowder, P.G.Carey, P.M.Smith, R.S.Sposili, H.S.Cho, and J.S.Im. Low temperature single crystal si tfts fabricated on si–filmw processed via sequential lateral solidification. In *IEEE Electron Device Letters, Vol. 19, No. 8*, pages 306–308, 1986.

[18] G. McFarland and M. Flynn. Limits of scaling mosfets. CSL TR-95-62, Stanford University, November 1995.

[19] M.Rodder and S.Aur. Utilization of plasma hydrogenization in stacked SRAMs with pli-Si PMOSFETs and bulk Si NMOS FETs. In *IEEE Electron Device Letters, Vol. 12*, pages 233–235, 1999.

[20] M.W.Geis, D.C.Flanders, D.A. Antoniadis, and H.I.Smith. Crystalline silicon on insulators by graphoepitaxy. In *IEDM Technical Digest*, pages 210–212, 1979.

[21] Ozcan Ozturk, Feng Wang, Mahmut T. Kandemir, and Yuan Xie. Optimal topology exploration for application-specific 3d architectures. In *ASP-DAC*, pages 390–395, 2006.

[22] S. Pae, T.-C. Su, J.P. Denton, and G/W. Neudeck. Multiple layers of silicon-on-insulator islands fabrication by selective epitaxial growth. In *IEEE Electron Device Letters, Vol. 20, No. 5*, 1999.

[23] S. Palacharla, N. P. Jouppi, and J. E. Smith. Complexity-effective superscalar processors. In *Proceedings of the 24th Annual International Symposium on Computer Architecture*, pages 206–218, June 1997.

[24] P.Wilkerson, A.Raman, and M.Turowski. Fast, automated thermal simulation for three-dimensional integrated circuits. In *Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Circuits, Itherm*, 2004.

[25] G. Reinman and N. P. Jouppi. Cacti 2.0 beta. In *In http ://www.research.digital.com/wrl//people/jouppi/CACTI.html, 1999*.

[26] P. Shivakumar and Norman P. Jouppi. Cacti 3.0: An integrated cache timing, power, and area model. In *Technical Report*, 2001.

[27] S.Kawamura, N.Sasaki, T.Iwai, M.Nakano, and M.Takagi. Three-dimensional cmps ics fabricated using beam recrystallization. In *IEEE Electron Devices Vol. Ed. 4*, pages 366–369, 1983.

[28] E. Sprangle and D. Carmean. Increasing processor performance by implementing deeper pipelines. In

*29th Annual International Symposium on Computer Architecture*, 2002.

[29] T.Kunio, K.Oyama, Y.Hayashi, and M.Morimoto. Three dimensional ics, having four stacked active decide layers. In *IEDM Technical Digest*, pages 837–840, 1989.

[30] Y. Tsai, Y. Xie, N. Vijaykrishnan, and M. Irwin. Three-dimensional cache design exploration using 3DCacti. In *International Conference on Computer Design*, October 2005.

[31] S. Wilton and N. Jouppi. Cacti: An enhanced cache access and cycle time model. In *IEEE Journal of Solid-State Circuits*, May 1996.

[32] Y.Akasaka and T. Nishimura. Concept and basic technologies for 3d ic structure. In *IEDM Technical Digest*, pages 488–491, 1986.