3D-NonFAR: Three-Dimensional Non-Volatile FPGA ARchitecture Using Phase Change Memory *

Yibo Chen, Jishen Zhao, Yuan Xie
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802
{yxc236, juz138, yuanxie}@cse.psu.edu

ABSTRACT

Memories play a key role in FGPAs in the forms of both programming bits and embedded memory blocks. FPGAs using non-volatile memories have been the focus of attention with zero boot-up delay, real-time reconfigurability, and superior energy efficiency. This paper presents a novel three-dimensional (3D) non-volatile FPGA architecture (3D-NonFAR) using phase change memory (PCM) and 3D die stacking techniques. Basic structures in a conventional FPGA architecture are renovated with PCM, and components are repartitioned and reorganized in 3D-NonFAR to allow an efficient 3D integration of PCM elements. 3D-NonFAR not only preserves the advantages of existing non-volatile FPGAs, but also provides high integration density, high performance, and bit-level programmability, which enable PCM as a universal memory replacement in FPGAs. Evaluation results show that 3D-NonFAR has smaller footprint, higher performance, and lower power consumption compared with other FPGA counterparts.

Categories and Subject Descriptors

B.7.1 [**Integrated Circuits**]: Types and Design Styles — *Gate arrays*

General Terms

Design

Keywords

3D IC, Non-Volatile FPGA, Phase-Change Memory

1. INTRODUCTION

Field Programmable Gate Arrays (FPGAs) have become a viable alternative to custom Integrated Circuits (ICs) by providing flexible computing platforms with short time-to-market and low costs. In an FPGA-based system, a design is mapped onto an array of reconfigurable logic blocks, which are connected by reprogrammable interconnections composed of wire segments and switch boxes. The configuration and programming bits are stored in SRAM cells and they are usually loaded from off-chip ROM or flash at boot-up.

While SRAM-based FPGA suffer from long configuration-loading time and excessive leakage power during stand-by, FPGAs using non-volatile memories (NVM) have emerged as a promising alternative [1]. Non-volatile memory eliminates the necessity of loading configuration from off-chip storage, for it preserves the configuration information stored on-chip while powered off, allowing the devices to immediately run at power-up. In addition, the immunity to soft errors also makes such NVM-based FPGA attractive for mission-critical aerospace applications [1]. Consequently, FPGAs using flash to store configuration bits have already been in the market by Actel and Lattice [1]. However, low logic density, inadequate performance, and the lack of bit-level programmability prevent flash memory from being the universal memory replacement in FPGAs. Recently, new non-volatile memory technologies including magnetic RAM (MRAM), Ferroelectric RAM (FeRAM), and phase change RAM (PCRAM) have been intensively explored [2]. These candidates generally provides high logic density and moderate to high performance compared with existing technologies. However, the manufacture of these new memories usually requires new materials and separate processes which complicate the fabrication of FPGAs. A low-cost integration technique is thus needed to incorporate these new non-volatile memories in conventional FPGAs.

To provide the required reconfigurable functionality, FPGAs provide a large amount of programmable interconnect resources in the form of wire segments, switches, and signal repeaters. Typically, the delay of FPGAs is dominated by these interconnects. Reducing the lengths of interconnects can lead to significant improvements in the performance of FPGAs.

As an emerging technology for integration, three dimensional (3D) ICs can increase the performance, functionality, and logic density of integrated systems. Recently, a number of publications has proposed novel 3D architectures and physical design techniques leading to better performance than that of existing planar FPGAs [3,4]. However, research work on using non-volatile memories in the 3D FPGA context is still in its infancy.

*This work was supported in part by NSF 0702617, 0916887, 0903432, and SRC grants.

ISLPED'10, August 18–20, 2010, Austin, Texas, USA.

Table 1: comparison of different memory technologies [5]

	SRAM	DRAM	NAND Flash	PCM
Cell size	$> 100F^2$	$6 - 8F^2$	$4 - 6F^2$	$4 - 20F^2$
Read latency	$\sim 10ns$	$\sim 10ns$	$5\mu s - 50\mu s$	$10 - 100ns$
Write latency	$\sim 10ns$	$\sim 10ns$	$2 - 3ms$	$100 - 400ns$
Access power	low	low	high	moderate
Standby power	leakage	refresh	zero	zero
Bit-level Alterability	Yes	Yes	No	Yes
Soft-error Immunity	No	No	Yes	Yes

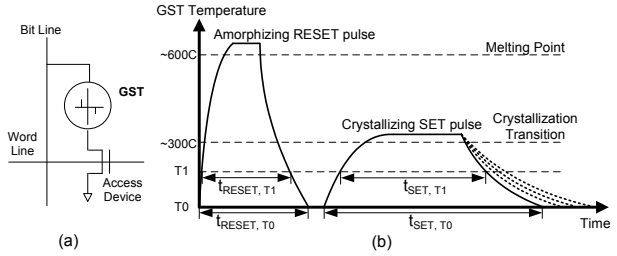

Figure 1: (a) The basic structure of PCM cell. (b) Write access to a PCM cell.

This paper presents a novel 3D non-volatile FPGA architecture (3D-NonFAR) based on the emerging phase change memory (PCM). With PCM's high performance, excellent scalability, and high density, we employ PCM as the universal memory replacement in FPGAs. Together with 3D integration using die stacking, the proposed 3D-NonFAR architecture exhibits smaller footprint, higher performance, and lower power consumption. We summarize the contributions of this paper as follows.

- We renovate basic structures of FPGAs including look-up table (LUT), switching boxes and block RAMs using PCM.
- We propose a novel scheme for resource partition and layer assignment in 3D-NonFAR, so that the chip footprint and the usage of inter-layer connections are minimized.
- We evaluate the proposed 3D-NonFAR in terms of logic density, timing and power, and compare the results with other 3D FPGA counterparts.

2. PRELIMINARIES ON PCM AND 3D INTEGRATION

Since PCM and 3D integration technology are still in a speculative state, researchers have made several different manufacturing and design decisions. This section surveys the latest research progresses on PCM and 3D integration, and derives corresponding device and technology choices for this paper.

2.1 PCM Technology

Phase change memory (PCM) technology is based on a chalcogenide alloy (typically, $Ge_2Sb_2Te_5$, GST) similar to those commonly used in optical storage (compact discs). Per the comparison with present memory technologies listed in Table 1, PCM is recognized as one of the most promising candidates for universal memory replacement in near future [5, 6].

The basic structure of a PCM cell consists of a GST and an access transistor, as shown in Fig. 1-(a). The data storage capability of a PCM cell is based on the property of the GST material to reversibly change between an amorphous and a crystalline phase when stimulated with adequate thermal pulses. Fig. 1-(b) shows the thermal pulses required for the PCM write operations, which are realized by applying electrical current pulses to the PCM cell. The GST material is programmed either to a high-resistance amorphous state (RESET to logic 0) or to a low-resistance crystalline state (SET to logic 1). The ratio between the high- and low-resistance is usually on the order of 10^2.

Multi-level cells (MLC) store multiple bits by programming the cell to produce intermediate resistances [7]. As shown in Fig. 1-(b), smaller current slopes (i.e., slow ramp down) produce lower resistances and larger slopes (i.e., fast ramp down) produce higher resistances. Varying slopes induce partial phase transitions and/or change the size and shape of the amorphous material produced at the contact area, leading to resistance values between those of fully amorphous or fully crystalline chalcogenide. Typically, MLC cells are constrained to have two bits per cell due to the difficulty of differentiating between a large number of resistances.

2.2 3D Integration Principles

With the continuous technology scaling, interconnect has emerged as the dominant source of circuit delay and power consumption. Three-dimensional (3D) ICs have recently recognized as a promising means to mitigate the interconnect-related problems [8, 9]. Several 3D integration technologies have been explored recently, including wire bonding, microbump, contactless (capacitive or inductive), and through-silicon-via (TSV) vertical interconnects [8]. TSV-based 3D integration has the potential to offer the greatest vertical interconnect density, and therefore is the most promising one among all the vertical interconnect technologies. In 3D IC chips that are based on TSV technology, multiple active device layers are stacked together (through wafer stacking or die stacking) with direct vertical TSV interconnects [9].

3D ICs offer a number of advantages over traditional two-dimensional (2D) design, including: (1) Higher packing density and smaller footprint; (2) Shorter global interconnect due to the short length of TSVs and the flexibility of vertical routing; (3) Higher performance because of the interconnect wire length reduction and bandwidth improvement; (4) Lower interconnect power consumption due to reduced interconnect capacitance; (5) Support of heterogenous integration: each single die can have different technologies.

The stacking of multiple active layers in 3D design leads to even higher power densities than for the 2D counterpart, thereby exacerbating the thermal and power problems, which have to be addressed when new 3D architectures are proposed.

3. 3D-NONFAR: PCM-BASED 3D FPGA ARCHITECTURE

In this section we propose the novel 3D non-volatile FPGA architecture (3D-NonFAR), including the renovation of basic

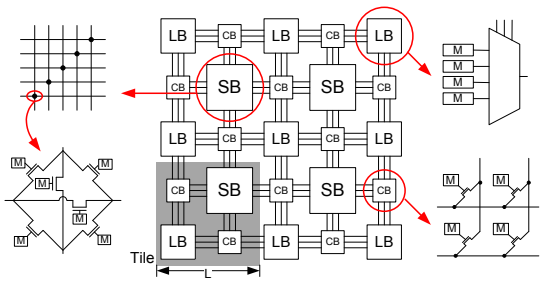

Figure 2: Classical SRAM-based FPGA architecture. (LB: Logic blocks; SB: Switching blocks; CB: Connection blocks; M: Memory elements)

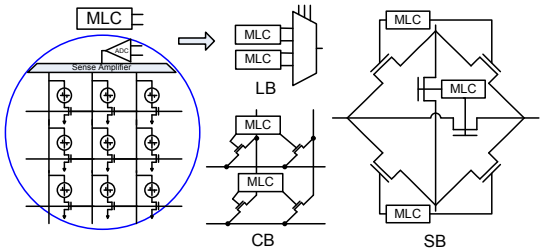

Figure 3: Renovated FPGA basic structures with PCM MLC cells.

FPGA structures, as well as the layer partition and logic density evaluation for the 3D die stacking.

3.1 PCM-Based FPGA Basic Structures

Firstly, we choose Xilinx Virtex-4 island-based FPGA [10] as the baseline architecture for our study. The baseline fabric, as shown in Fig. 2, consists of a 2-D array of logic blocks (LBs) that can be interconnected via programmable routing. Each LB contains four slices, each consisting of two four-input lookup tables (LUTs). The programmable routing comprises interconnect segments that can be connected to the LBs via connection boxes (CBs) and to each other via switching boxes (SBs).

Programming bits are used to implement logic function in LUTs and to determine logic connectivity in CBs and SBs. Conventionally, these bits are loaded from off-chip at power-up and stored in SRAM cells. Modern FPGAs also have embedded RAM blocks for high-bandwidth data exchange during the computation. With the PCM technology, the SRAM cells in FPGAs can be replaced with PCM cells as follows, and the logic density can be significantly improved according to the cell size data listed in Table 1.

- Configuration bits in LBs, CBs, and SBs. Writing to these bits only happens at configuration time. During FPGA run time, these bits are only read. Thus the write speed for these bits is not an issue, and they can be implemented using PCM MLC cells with a slightly slower write speed than that of SLC cells. The logic density improvement per bit of PCM MLC against SRAM is about 16× [7]. The renovated basic FPGA structures with PCM MLC cells is shown in Fig. 3.

- Embedded RAM blocks. The read/write speed of RAM blocks is critical for the performance of FPGAs, and they can be implemented using PCM SLC cells. From Table 1, PCM SLC write latency is 10×-40× larger than that SRAM. However, through a set of write optimizations such as write coalescing [6] and aggressive word-line/bit-line cutting [5], PCM can achieve about the same performance as DRAM ([6] reported 1.2x slower than DRAM). Simulation results [5] show that the logic density improvement per bit against SRAM is about 10×.

3.2 3D PCM-Based FPGA with Die Stacking

In addition to the high-performance reconfigurable logic fabric, modern FPGAs also have many built-in system-level blocks, including microprocessor cores, DSP cores, I/O transceivers, etc. These features allow logic designers to build the highest levels of performance and functionality into their FPGA-based systems. In order to efficiently integrate these system-level blocks in the new architecture, their characteristics need to be accurately modeled.

The areas of components including system-level blocks are estimated using the models presented in previous studies [3, 11]. An area breakdown of Xilinx Virtex-4 Platform FPGAs based on SRAM is depicted in Fig. 4, where SRAM occupies about 40% of the total area of planar FPGAs. Replacing the SRAM with high-density PCM cells can significantly improve the logic density and reduce chip area.

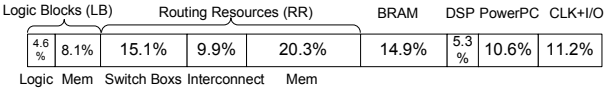

Figure 4: Area breakdown of Xilinx Virtex-4 Platform FPGAs.

Heterogeneous three-dimensional (3D) die stacking is an effe ctive way to further improve the logic density and reduce the chip footprint. The stacking for FPGA begins with partitioning the resources and assigning them to different layers. We use a simulated-annealing based layer assignment algorithm similar to [12], to minimize the following cost function:

$$cost = \alpha * (A + dev(A)) + \beta * n_{TSV},$$
$$dev(A) = \sum_{i=1}^{N} |A(i) - A_{avg}| \quad (1)$$

where (A_i) is the area of the die in layer i, $A = max(A(i)$ is the footprint of the chip, $A_{avg} = \sum_{i=1}^{N} A(i)/N$ is the average area per layer, and n_{TSV} is the number of TSVs. To reduce the integration cost and improve the chip performance, the following constraints are also set for the layer assignment algorithm:

- Basic FPGA infrastructures are preserved so that existing FPGA CAD tools can still be used for the new architecture to maintain low integration cost;

- The partition and integration of non-configurable system-level blocks (processor cores, DSP blocks, etc) are fully addressed;

- All non-volatile memory elements are aggregated in one single layer to reduce manufacture cost;

With all the specifications listed above, the layer assignment algorithm produces *3D-NonFAR*, an novel 3D non-volatile FPGA architecture with two-layer die stacking, as demonstrated in Fig. 5. In 3D-NonFAR, all CBs and SBs, as well as the memory elements in LBs and the embedded

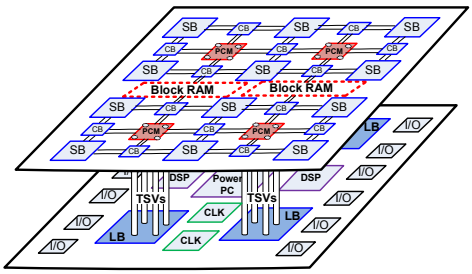

Figure 5: 3D-NonFAR: The proposed 3D non-volatile FPGA architecture with two-layer die stacking and TSVs.

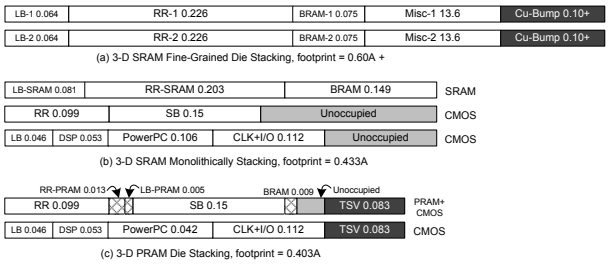

Figure 6: Chip footprint comparison between different 3D FPGA stacking scenarios. The numbers in each block represent the relative area compared to the area of the baseline 2D FPGA.(A, area of baseline 2D FPGA; RR, routing resource; LB-SRAM, configuration memory cells in LB; RR-SRAM, configuration memory cells in RR; ST, switch transistor).

block memories are put in the upper layer, while the LBs are located in the lower layer and TSVs are used to connect the LBs to the corresponding memory elements upward. Since all the configurable routing tracks are in the upper layer, no TSV is needed for routing. The non-configurable system-level blocks and the clock network are also located in the same layer as the LBs, thus no TSV is needed to delivery clock signals to the flip-flops in LBs.

We compare the logic density improvement of 3D-NonFAR with other two 3D SRAM-based FPGA stacking scenarios, namely *3D-FineGrain* [3] in which 3D (6-way) switching blocks are used and all the resources evenly distributed between layers in fine granularity, and *3D-Mono* [4] in which all SRAM cells are moved to a separate layer and monolithical stacking is used so that there is no overhead on inter-layer connections[1]. We conduct the comparison on Virtex-4 XC4VFX60 Platform FPGA, which is in the same technology node (*90nm*) as the latest PCM cells in literature [7]. The pitch of Cu bumps used in 3D-FineGrain for inter-layer connection is 10 μm. Based on data reported in [13] we choose typical 3×3 μm^2 TSV pitch for the experiments (TSV size is $\sim$ 1 μm^2). The number of Cu bumps and TSVs are estimated according to the partition style of logic fabrics and the usage of non-configurable blocks. For 3D-FineGrain,

$$n_{bump} \simeq W * n_{SB}/2 \qquad (2)$$

[1]In monolithical stacking, electronic components and their connections (wiring) are lithographically built in layers on a single wafer, hence no TSV is needed. Applications of this method are currently limited because creating normal transistors requires enough heat to destroy any existing wiring.

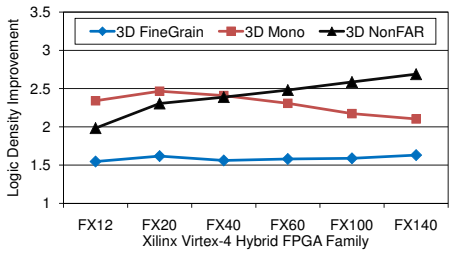

Figure 7: Logic density improvement of three 3D FPGA stacking styles. By improvement, we mean the ratio of footprint in baseline 2D FPGA to that in 3D FPGA.

where W is the channel width of routing fabrics and $n_S B$ is the number of SBs. For 3D-NonFAR,

$$n_{TSV} \simeq (4+4+1) * n_{LUT4} + 2 * n_{GlobalWire} \qquad (3)$$

where n_{LUT} is the number of 4-input LUTs, and $n_{GlobalWire}$ is the number of global interconnects that go across layers.

Fig. 6-(a) shows the case of 3D-FineGrain, in which 3D switching blocks incurs massive usage of Cu bumps and the footprint reduction is limited to be around 40%. In Fig. 6-(b), 3D-Mono leads to unbalanced layer partitions with a separate SRAM layer, and the footprint reduction is about 56% with three-layer stacking. Fig. 6-(c) shows that in 3D-NonFAR, PCM cells occupy much smaller area than SRAM cells, so the upper layer can accommodate more blocks, yielding the most area-efficient stacking with a footprint reduction of near 60%. The logic density improvement across the whole Xilinx XC4VFX FPGA Family is depicted in Fig. 7, which shows that 3D-NonFAR is more favorable in larger devices.

4. PERFORMANCE AND POWER EVALUATION OF 3D-NONFAR

In this section, we quantify the improvements in delay and power consumption of 3D-NonFAR. To accurately model the impact of integrating PCM as a universal memory replacement in FPGAs, we first evaluate the timing and power characteristics of memories with CACTI [14] and PCRAMsim [5], as well as the data from latest literature [7]. A set of evaluation results are listed in Table 2. We then proceed to evaluate the delay and power of the whole chip.

Table 2: characteristics of 32MB RAM blocks in different memory technologies at 90nm process

	SRAM	DRAM	PCM (SLC)	PCM (MLC)
Area (mm^2)	523.5	32.2	59.2	36.0
Read Latency (ns)	13.1	6.92	10.7	~15
Write Latency (ns)	13.1	6.92	151.8	285.7
Read Energy (nJ)	5.03	1.07	2.26	3.49
Write Energy (nJ)	2.59	0.55	4.78	7.17
Leakage Power (nW)	5.14	1.71	0.0091	0.0137

4.1 Delay Improvement in 3D-NonFAR

The path delay in FPGAs can be divided into logic block delay and interconnect delay. In 3D-NonFAR, the LUTs in logic blocks are implemented with PCM MLC cells, which are about 5% slower in read speed than that of conventional SRAM cells, as shown in Table 2. However, this overhead can be easily compensated by the reduction of interconnect

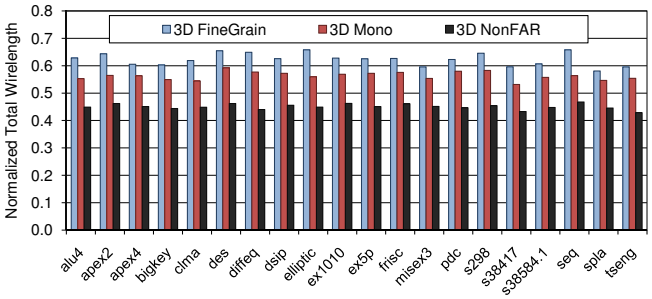

Figure 8: The total wirelength reduction by 3D FPGA stacking. For each benchmark, the wirelength values are normalized against the total wirelength of the baseline 2D FPGA.

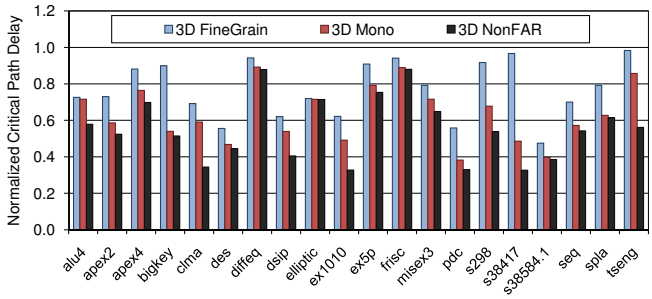

Figure 9: The critical path delay reduction by 3D FPGA stacking. For each benchmark, the delay values are normalized against the critical path delay of the baseline 2D FPGA.

delay which will be assessed later in this section. It also has to be mentioned that, although the write operation of PCM MLC cells is much slower, it will not affect the runtime performance of LUTs, because write to LUTs only occurs during the configuration time.

The block RAMs in 3D-NonFAR are implemented in PCM SLC cells. As aforementioned, they can achieve the same performance as that of SRAM blocks with aggressive optimizations [5], so we assume that no delay overhead is incurred by block RAMs.

The reduction of interconnect delay is quantified as follows. FPGAs have wire segments of different lengths that can be used to implement interconnections with different requirements. In Xilinx Virtex-4 FPGAs, there are four types of wire segments, which are with length 1 (Single), 3 (Hex-3), 6 (Hex-6) and 24 (Global) (The unit for the wire length is the tile width L). The RC delay of each type of wire segments can be computed from the Elmore delay model in [4]. For 3D FPGA stacking, the unit wire length L is reduced by a logic density scaling factor η. While we assume that all of Single, Hex-3, and Hex-6 wires reside in the upper layer, Global wires can go across layers with TSVs for interconnect reduction. For TSVs on Global wires, we use the resistance and capacitance values of $43m\Omega$ and $40fF$ respectively for delay calculation. We feed the improved unit length to the Elmore model [4] and compute the delay improvement for each type of wire segments. The results are listed in Table 3.

Table 3: Relative Delay Improvement of Different Types of Interconnect wire segments

Interconnect Type	Single	HEX-3	HEX-6	Global
3D-FineGrain ($\eta = 0.62$)	1.28	1.51	1.61	1.63
3D-Mono ($\eta = 0.44$)	1.39	1.60	1.73	1.87
3D-NonFAR ($\eta = 0.41$)	1.62	2.05	2.16	2.58

We then evaluate the improvement of overall system performance for 3D-NonFAR as well as 3D-FineGrain [3] and 3D-Mono [4] architectures. Since both 3D-NonFAR and 3D-Mono maintain the basic infrastructure of conventional 2D FPGA, they can be evaluated using the commonly-used VPR toolset [15]. 3D-FineGrain can be modeled within TPR [16], which is an variant of VPR and supports the modeling of fine-grained 3D FPGA integration. The evaluation is performed in the following procedure:

1. We modify VPR/TPR so that they recognize the cluster structure in Virtex-4 FPGAs and generate the correct routing graph.

2. We then modify the architecture models in VPR/TPR to reflect the revised delay values of logic blocks and wire segments, according to the parameters scaled from Table 2 and Table 3. The size of buffers inserted in long interconnects are also re-optimized with the new delay parameters;

3. We place and route the 20 largest MCNC benchmark circuits with the modified VPR/TPR, and quantify the relative improvement against the baseline 2D FPGA.

Fig. 8 depicts the total wire length reduction brought by 3D FPGA stacking. The average reductions against the baseline 2D FPGA for 3D-FineGrain, 3D-Mono and 3D-NonFAR are 37.7%, 43.7% and 54.9%, respectively. Fig. 9 depicts the critical path delay reduction. The average reductions for 3D-FineGrain, 3D-Mono and 3D-NonFAR are 22.9%, 36.5% and 44.9%, respectively. We can see that the total wire length is nearly uniformly reduced across all the benchmarks, while the critical path delay improvement largely depends on the circuit structures.

4.2 Reduction of Power Consumption in 3D-NonFAR

There are two major components of power consumption in FPGAs: static power, and dynamic power. Although the dynamic component is still dominating in FPGAs in 90nm process technology, static power occupies a significant portion in the total power consumption [17]. We thus have to quantify the reduction in both dynamic and static power in 3D-NonFAR.

The breakdown of dynamic power in FPGAs is shown as:

$$\begin{aligned} P_{dyn} &= P_{logic,dyn} + P_{mem,dyn} + P_{net,dyn} + P_{clk,dyn} \\ &= P_{logic,dyn} + P_{mem,dyn} \\ &\quad + \psi V_{DD}^2 f_{net} \sum_{all\ nets} C_{net} + C_{clk} V_{DD}^2 f_{clk}, \end{aligned} \quad (4)$$

where $P_{logic,dyn}$ is the dynamic energy consumed in the logic circuits (including the non-configurable blocks, e.g. DPSs), $P_{mem,dyn}$ in the memory elements, $P_{int,dyn}$ in the interconnect wires, and $P_{clk,dyn}$ in the clock network. In the 2D baseline FPGA, the ratio between $P_{logic,dyn}$, $P_{mem,dyn}$, $P_{int,dyn}$, and $P_{clk,dyn}$ is estimated to be 0.3:0.15:0.4:0.15 using Xilinx X-Power tool [18]. During the transition from 2D FPGAs to

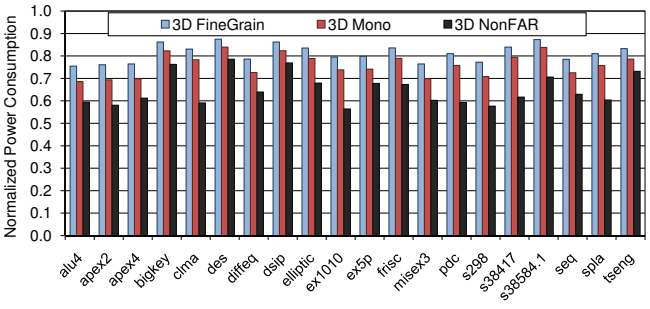

Figure 10: The total power reduction by 3D FPGA stacking. For each benchmark, the power values are normalized against the total power of the baseline 2D FPGA.

3D FPGAs, $P_{logic,dyn}$ is assumed to be unaffected, while the change of $P_{mem,dyn}$ can be estimated with the parameters in Table 2. According to Table 2, the write energy for PCM cells is much higher than that of SRAM cells. However, the total write energy can be reduced using narrow rows and partial writes as presented in [6].

$P_{int,dyn}$ and $P_{clk,dyn}$ are determined by the equivalent capacitances of the signal net and the clock network (C_{net} and C_{clk}, respectively). While other parameters contributing to $P_{int,dyn}$ and $P_{clk,dyn}$ remain unaffected, C_{net} and C_{clk} are scaled down with the logic density scaling factor η, leading to the major reduction in dynamic power. Note that for Global wires involving TSVs, the capacitance of TSVs are included in the re-calculation of C_{net}.

As for the leakage power, the portion consumed on logic circuits in 3D-NonFAR is the same as that in the baseline 2D FPGA, while the portion of memory cells is significantly reduced, due to the zero leakage of PCM cells. This can be updated according to Table 2.

To evaluate the total power reduction of 3D-NonFAR, we assume the ratio between the total dynamic power and static power consumed in baseline 2D FPGAs is 0.8:0.2 [17]. The power numbers for the logic circuits including LBs, DPSs, PowerPC cores, and IO blocks are provided by Xilinx XPower [18], while the values for interconnects and clock network are estimated using the breakdown ratios. The total power for each of the 20 MCNC benchmark circuit is estimated according to the resource usage, and the relative power reduction against the baseline 2D FPGA for 3D stacking is depicted in Fig. 10, where the average reductions for 3D-FineGrain, 3D-Mono, and 3D-NonFAR are 18.7%, 24.0% and 35.1%, respectively.

5. CONCLUSION

In this paper we presented a novel three-dimensional non-volatile FPGA architecture (3D-NonFAR) using phase change memory (PCM) and 3D die stacking techniques. We renovated the basic structures in FPGA with non-volatile PCM cells, and partitioned the resources to build up a 2-layer footprint-efficient stacking. We then quantified the improvement on logic density, delay and power consumption. We also addressed the consequent thermal and reliability issues in the new architecture. Evaluation results suggest that the proposed 3D-NonFAR is a promising alternative for future FPGA technology innovation.

6. REFERENCES

[1] K. Han, N. Chan, and et al. A novel flash-based FPGA technology with deep trench isolation. In *IEEE Non-Volatile Memory Workshop*, 2007.

[2] R. Bez and A. Pirovano. Non-volatile memory technologies: emerging concepts and new materials. *Materials Science in Semiconductor Processing*, 7(4-6):349 – 355, 2004.

[3] A. Rahman, S. Das, and et al. Wiring requirement and three-dimensional integration technology for field programmable gate arrays. *TVLSI*, 11(1):44–54, 2003.

[4] M. Lin, A. El Gamal, and et al. Performance benefits of monolithically stacked 3-D FPGA. *TCAD*, 26(2):216–229, 2007.

[5] X. Dong, N. Jouppi, and Y. Xie. PCRAMsim: System-level performance, energy, and area modeling for phase-change RAM. *ICCAD*, 2009.

[6] B. C. Lee, E. Ipek, and et al. Architecting phase change memory as a scalable DRAM alternative. In *ISCA*, 2009.

[7] F. Bedeschi, R. Fackenthal, and et al. A bipolar-selected phase change memory featuring multi-level cell storage. *JSSC*, 44(1):217–227, 2009.

[8] W. R. Davis, J. Wilson, and et al. Demystifying 3D ICs: the pros and cons of going vertical. *IEEE Design & Test of Computers*, 22(6):498–510, 2005.

[9] Y. Xie, G. H. Loh, and et al. Design space exploration for 3D architectures. *J. Emerg. Technol. Comput. Syst.*, 2(2):65–103, 2006.

[10] Xilinx. Virtex-4 FPGA data sheets. `http://www.xilinx.com/support/documentation/virtex-4_data_sheets.htm`.

[11] C. H. Ho, W. Leong, and et al. Virtual embedded blocks: A methodology for evaluating embedded elements in FPGAs. In *FCCM*, 2006.

[12] P.H. Shiu, R. Ravichandran, S. Easwar, and S.K. Lim. Multi-layer floorplanning for reliable system-on-package. 2004.

[13] Subhash Gupta, Mark Hilbert, and et al. Techniques for producing 3D ICs with high-density interconnect. `http://www.tezzaron.com/about/papers/ieee%20vmic%202004%20finalsecure.pdf`, 2004.

[14] S. Thoziyoor, N. Muralimanohar, and N. P. Jouppi. CACTI 5.1. Technical Report HPL-2008-20, HP Laboratories, 2008.

[15] V. Betz and J. Rose. VPR: A new packing, placement and routing tool for FPGA research. In *Field-Programmable Logic Applications*, 1997.

[16] C. Ababei, P. Maidee, and K. Bazargan. Exploring potential benefits of 3D FPGA integration. In *FPL*, 2004.

[17] A. Telikepalli. Power-performance inflection at 90 nm process node FPGAs in focus. In *Chip Design Magazine*, April 2008.

[18] Xilinx. XPower estimator. `http://www.xilinx.com/products/design_tools/logic_design/verification/xpower.htm`.