Physical Design of 3D FPGAs Embedded with Micro-channel-based Fluidic Cooling

Zhiyuan Yang and Ankur Srivastava
University of Maryland, College Park, MD, USA
{zyyang, ankurs}@umd.edu

ABSTRACT

Through Silicon Via (TSV) based 3D integration technology is a promising technology to increase the performance of FPGAs by achieving shorter global wire-length and higher logic density. However, 3D FPGAs also suffer from severe thermal problems due to the increase in power density and thermal resistance. Moreover, past work has shown that leakage power can account for 40% of the total power at current technology nodes and leakage power increases non-linearly with temperature. This intensifies the thermal problem in 3D FPGAs and more aggressive cooling methods such as micro-channel based fluidic cooling are required to fully exploit their benefits. The interaction between micro-channel heat sink design and the performance of a 3D FPGA is very complicated and a comprehensive approach is required to identify the optimal design of 3D FPGAs subject to thermo-electrical constraints. In this work, we propose an analysis framework for 3D FPGAs embedded with micro-channel-based fluidic cooling to study the impact of channel density on cooling and performance. According to our simulation results, we provide guidelines for designing 3D FPGAs embedded with micro-channel cooling and identify the optimal design for each benchmark. Compared to naive 3D FPGA designs which use fixed thermal heat sink, the optimal design identified using our framework can improve the operating frequency and energy efficiency by up to 80.3% and 124.0%.

1. INTRODUCTION AND MOTIVATION

With the continued dimensional scaling, the non-recurring engineering cost in the design of cell-based Application Specific Integrated Circuits (ASICs) escalates enormously. Compared to ASICs, FPGAs have the advantage of higher reusability, simpler design cycle and faster time-to-market. On the other hand, FPGAs suffer from tremendous programming overheads. According to [12], programmable elements can account for 90% total footprint area, 80% total path delay and large portion of power consumption. These overheads enlarge the performance gap between FPGAs and ASICs [9] and become the main limitation to the growth of FPGA market.

3D integration technology provides an opportunity to increase the performance of FPGAs. This emerging technology stacks multiple dies on top of each other and uses

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FPGA'16, February 21-23, 2016, Monterey, CA, USA

© 2016 ACM. ISBN 978-1-4503-3856-1/16/02...\$15.00

DOI: <http://dx.doi.org/10.1145/2847263.2847275>

Figure 1: Trend of leakage power with increasing temperature. (Data are normalized to the leakage power at 85°C)

Through Silicon Vias (TSVs) for inter-die connection which achieves shorter global wire-length and higher transistor density. However, 3D integration is not a panacea and it introduces a number of new challenges: (1) relatively larger TSV dimension will cause the increase in footprint area, (2) switch boxes that provide inter-layer connection will use more transistors resulting in extra power and delay, and (3) stacked structure will exacerbate thermal problem by increasing both the power density and thermal resistance.

Recently, several significant researches have been done to exploit 3D integration in FPGAs. Authors of [6] and [15] studied the design of switch box topology in 3D FPGAs based on the fact that the number of TSVs is limited in 3D ICs. Pangracious et al.[14] proposed to use tree-based interconnect topology to build 3D FPGAs in order to reduce the area overhead and improve logic density. Instead of wafer-stacked 3D FPGAs, people also studied monolithic 3D FPGAs [12, 11] which, according to the results, have better performance. Meanwhile, placement and routing (P&R) tools for 3D FPGAs have also been developed, among which TPR [3] and 3D-Meander [24] are the most popular ones. Unfortunately, despite of the work that has been done, thermal issues are almost unexplored in 3D FPGAs. Conventionally, 2D FPGAs dissipate a small amount of power due to low frequency and low logic density. Therefore, they can be sufficiently cooled by air-based heat sink. With 3D integration technology, power density of 3D FPGAs will increase due to the improvement of frequency and logic density while the heat removal capacity of conventional air-based cooling does not improve accordingly. This will lead to the rise of on-chip temperature. Authors of [6] characterized the temperature of a type of 3D FPGAs based on Virtex-4 architecture and discovered a 2.5x increase in peak temperature when the number of layers increases from one to four. If the on-chip temperature exceeds over the temperature limit, operating frequency should be scaled down to make the design ther-

Figure 2: Illustration of the architecture of an island-style FPGA

mally feasible, which results in performance degradation in 3D FPGAs.

Thermal problems can be further exacerbated when the leakage power is taken into consideration. With the geometric and supply voltage scaling, leakage power becomes the primary contributor to the total power dissipation in FPGAs. Several researches [4, 26] demonstrate that leakage power can account for as much as 40% of the total power in FPGAs at sub-100nm technology nodes. Moreover, leakage power increases non-linearly with temperature [20] (as illustrated in Figure 1). The positive feedback between temperature and leakage power will lead to thermal runaway if 3D FPGAs are not sufficiently cooled. In order to fix the thermal problem and fully exploit the benefits of 3D FPGAs, more aggressive cooling approaches, such as micro-channel-based fluidic cooling, should be applied.

Micro-Channel-Based Fluidic Cooling (MC Cooling) comprises micro-channels etched into silicon substrate of each layer in 3D ICs. Fluid coolant (usually de-ionized water) is pumped into micro-channels and forms a distributed heat sink. This kind of micro-fluidic cooling is more efficient than air-cooling, because it can directly take heat away from adjacent regions instead of merely removing heat from top of the chip. According to [8, 27], single-phase micro-fluidic cooling can provide heat removal capacity as high as $700W/cm^2$ and two-phase cooling is even more efficient [7]. Recently, researchers built a micro-channel cooling system in the Altera FPGA with 28nm technology node and achieved much lower chip temperature compared to the air cooling [18]. Despite of these benefits, MC Cooling comes with other overheads: (1) TSVs cannot be built through micro-channels which results in a trade-off between micro-channel density and TSV-based vertical bandwidth; (2) extra energy is required to pump the coolant in order to cool the chip even though the pumping power is quite low according to the previous work [28, 16, 13]. For years, the application of micro-channel cooling in 3D ASICs has been widely studied [17, 22] and a 2D FPGA with liquid-cooling was reported recently [18]. However, the design-time optimization of 3D FPGAs with micro-channel cooling has not been investigated.

The scenario of applying MC Cooling in 3D FPGAs is far more complicated than that for 3D ASICs. On one hand, reducing micro-channel density will lead to worsening cooling. And increasing temperature may make 3D FPGAs thermally infeasible. On the other hand, due to the trade-off between channel density and vertical bandwidth, decreasing channel density leads to the increase of vertical bandwidth which may not necessarily reduce the delay of 3D FPGAs. This

is because that although larger vertical bandwidth will improve the latency of 3D nets, it will also cause congestion in 2D routing channels around TSVs due to the restriction of routing resources. This may increase the delay of the whole system in some cases. As for power dissipation, although increasing vertical bandwidth will possibly reduce the use of switch boxes, the average number of transistors per switch box will increase since more transistors are needed to support vertical connection and this will cause the rise of leakage power of a single switch box. Moreover, the dynamic power is determined by the operating frequency which is bounded by the routing delay. Hence the impact of channel density on power dissipation of 3D FPGAs is also nonmonotonic. Therefore, the relationship between cooling, operating frequency and power is very complicated and we cannot find an optimal design based on a simple model.

A comprehensive analysis approach is required to study the impact of the design of micro-channel heat sink on the 3D FPGA performance. Guidelines for designing 3D FPGAs embedded with micro-channel-based fluidic cooling are also necessary.

The contribution of our work is as follows:

- (1) We propose an analysis framework to study the impact of micro-channel density on performance and power of 3D FPGAs.
- (2) According to the simulation results of our framework, we provide guidelines for designing 3D FPGAs embedded with MC Cooling.
- (3) Our framework sweeps different micro-channel densities and identifies the optimal physical design of micro-channel heat sink and 3D FPGAs. Compared to naive designs using fixed micro-channel density, our framework can improve the operating frequency and energy efficiency by up to 80.3% and 124%, respectively.

2. MODELING OF FPGAS

FPGAs are a kind of integrated circuits that can be programmed by customers. Based on different types of global routing architecture, FPGAs can be categorized into hierarchical style and island style [9]. In island-style FPGAs, Configurable Logic Blocks (CLBs) and I/O blocks are arranged in a two dimensional mesh and routing fabrics (including Connection Boxes (CBs), Switch Boxes (SBs) and routing channels) are evenly distributed throughout the mesh. Figure 2 illustrates the architecture of an island-style FPGA. As illustrated by the figure, an island-style FPGA have a regular pattern and can be regarded as the construction of identical tiles [9]. Each tile contains a CLB with its adjacent routing fabrics as illustrated in Figure 2. Due to the regular pattern, island-style FPGAs exhibit a number of desirable properties including efficient connect between CLBs and routing tracks and high scalability. These properties make island-style the most commonly used architecture in modern SRAM-based FPGAs [2]. In this work we will use 2D island-style FPGAs as the baseline and stack several identical FPGA chips vertically to form a 3D FPGA.

2.1 3D FPGA Model

In our work, a 3D FPGA is modeled by vertically stacking a number of identical 2D island-style FPGA chips as illustrated in Figure 3. Some of the switch boxes in 2D FPGAs should be extended to support inter-layer connection and these switch boxes are called 3D switch box (3D-SB) while the others are 2D switch box (2D-SB). Inter-layer connection is realized with TSVs which are fabricated between two 3D-SBs placed on adjacent layers. In our model, we assume that each 3D-SB connects to identical number of TSVs. A primary concern about 3D FPGAs is the design of switch boxes. In this work, the so-called “subset-style” switch boxes [3,

Figure 3: Illustration of the architecture of a 3D stacked FPGA

Figure 4: Illustration of the subset topology of (a) 2D-SB and (b) 3D-SB

24] are employed while they can be replaced with any other advanced switch box architectures [6]. The topology of subset-style switch boxes is illustrated in Figure 4. In a switch box, an incoming routing track from one side can connect to routing tracks from other sides with the identical ID number. Therefore, the flexibility (F_s) of a 2D-SB equals three while a 3D-SB has $F_s = 5$ (including two TSVs connecting to the upper and lower layer respectively). Therefore, if both vertical and horizontal routing bandwidth in 3D FPGAs are equal to W_{route} , the number of transistors used in a 2D-SB is $6 \times W_{route}$ while the number of transistors in a 3D-SB is $15 \times W_{route}$. In practice, however, since TSVs occupy much more silicon area compared to 2D routing channels, we have to limit the number of TSVs connected by each 3D-SB (usually the number is smaller than the width of a 2D routing channel).

Another concern is the electrical characterization of TSVs. In this work, we use the following equations [19] to calculate the resistance and capacitance of a TSV:

$$R_{TSV} = \frac{\rho_m H_{TSV}}{\pi r_{via}^2} \quad (1)$$

$$C_{TSV} = \frac{2\pi\epsilon_r\epsilon_0 H_{TSV}}{\ln((r_{via} + t_{ox})/r_{via})} \quad (2)$$

In the equations, ρ_m is the resistivity of the metal filling in a TSV, r_{via} is the radius of the metal filling region in the TSV, t_{ox} is the sickness of the oxide layer around the TSV and ϵ_r is the relative permittivity of SiO_2 .

2.2 Area Model of 3D FPGAs

Calculation of temperature requires accurate modeling of 3D FPGA area. In this work, we follow the methodology introduced in [5] to determine the area of 3D FPGAs. According to [5], FPGA area is modeled as the number of “minimum width transistor areas”. The “minimum width transistor area” is the area of the layout of the smallest transistor, plus the minimum spacing to another transistor above it and to its right. In order to calculate the number of minimum width transistor areas, we first determine the number and size of transistors used in each programmable block (i.e. CLB, CB and SB). The schematic of each programmable block in our model is similar to the one described in [5]. For transistors

Figure 5: Illustration of a 3D stacked FPGA embedded with micro-channel cooling

with different sizes, their minimum width transistor area (MWTa) is calculated with:

$$\text{MWTa} = 0.5 + \frac{\text{Drive Strength}}{2 \times \text{DSMWT}} \quad (3)$$

where DSMWT stands for the “Drive Strength of Minimum Width Transistor”. Following this, we can calculate the number of minimum width transistor areas for each programmable block and the “real” area is computed by multiplying the number of minimum width transistor areas with the area value of a certain technology node. Note that up to this stage, TSVs are not considered yet. In our model, TSVs are assumed to be built between two 3D-SBs placed on adjacent layers. Therefore, the area of a 3D-SB should be added with the total silicon area occupied by TSVs connected to the 3D-SB.

In order to simplify the analysis, in this work, we extend the “tile” of an island-style FPGA as introduced in Section 2 in the context of 3D FPGAs. By doing this, a tile can be categorized into 3D-tile or 2D-tile based on the different types of switch boxes constituting the tile. This is illustrated in Figure 3. The area of a tile is the total area of all the programmable blocks included in the tile.

2.3 Characterizing The Impact of Micro-Channel Density In 3D FPGAs

The structure of 3D FPGAs embedded with micro-channel-based fluidic cooling is illustrated in Figure 5. Micro-channel structure is characterized by a series of physical parameters: channel width (W_{ch}), silicon thickness between channels (W_{wall}), channel height (H_{ch}) and channel length (L_{ch}). Usually, the length of the channel is equal to the length of the FPGA chip ($L_{ch} = L_{chip}$). TSVs of 3D FPGAs cannot be built through micro-channels resulting in a trade-off between vertical bandwidth and micro-channel density. In this work, we assume the silicon thickness between channels can only be equal to or be a multiple of channel width ($W_{wall} = \lambda W_{ch}$, $\lambda = 1, 2, 3, \dots$). Our work is going to study how the channel density affects cooling capacity and performance of 3D FPGAs.

Because a 3D FPGA is modeled by the array of tiles, we will capture the impact of micro-channel density on the distribution of different types of tiles in the 3D FPGA. In order to do this, we initially set 3D-tile as the default type of tile for all tiles. After this, we change some 3D-tiles to 2D-tiles such that micro-channels can be allocated, since micro-channels can only be allocated below the 2D-tiles due to their conflict with TSVs. It should be noted that, we assume a 2D-tile occupies the same area as a 3D-tile in a single layout even though they may require different areas. After mapping the micro-channel distribution to the 3D FPGA architecture, we can characterize different micro-channel densities by changing the distribution of 3D and 2D tiles in FPGAs. The distribution of different types of tiles will further influence the placement and routing in 3D FPGAs.

Figure 6: Flow chart of our analysis framework

3. ANALYSIS APPROACH FOR 3D FPGAS WITH MICRO-CHANNEL COOLING

In this section, we will introduce the method we used to study the impact of micro-channel density on the performance, power and energy efficiency of 3D FPGAs. In this work, we use the operating frequency to represent the performance. Energy efficiency is similar to the inverse of energy-delay-product (EDP) and is defined by the following equation:

$$E.E. = \text{Frequency}^2 / \text{Power} \quad (4)$$

Figure 6 illustrates the flow chart of the framework of our analysis methodology. The kernel of the framework is extended from TPR[3]. A given circuit (in blif format) is fed into T-Vpack (which is part of VPR[5], a P&R tool for 2D FPGA) and ACE2.0[10] (an activity estimation tool) to generate CLB-level netlist and transition density, respectively. Then, CLB-level netlist, FPGA architecture and distribution of micro-channels are fed into TPR for placing and routing. Meanwhile, TPR also calculates the final delay of the circuit based on its embedded delay model. Following this, we take the placement file, routing file as well as the delay and transition density as the inputs to our power model to calculate the power dissipation profile. The power profile is used to calculate temperature profile which is then fed back into power model to update the leakage power. This “power-thermal” loop is necessary because of the positive feedback between leakage power and temperature. The “power-thermal” loop stops when the update of temperature profile is negligible.

3.1 Background of TPR

TPR is one of the most popular P&R tools for 3D FPGAs. Originally, TPR can only support the 3D FPGA architecture with fully-vertical inter-connection. We extend the tool so that it can support different types of switch boxes (i.e. 2D-SB and 3D-SB). In this subsection, we will briefly introduce the principles of TPR. More detailed descriptions can be found in [3].

TPR takes the CLB-level netlist and the 3D FPGA architecture as inputs. In this work, we add micro-channel distribution as another input. In the first stage, the netlist is partitioned into different layers to minimize the total number of TSVs. Following this, TPR performs timing-driven partitioning-based placement successively from the top layer to the bottom layer. After the position of each CLB is determined, the tool uses Pathfinder Negotiated Congestion-delay algorithm [5] to assign routing elements to each net based on the distribution of routing resources. After routing is finished, the delay of the circuit is calculated based on Elmore Delay Model. Finally, TPR will output placement file, routing file and delay which will be used in the following stages.

3.2 Power Model

In order to compute the temperature and evaluate the energy efficiency of 3D FPGAs, we should first calculate the power dissipation profile. Since a 3D FPGA is modeled by an array of 3D and 2D tiles, we assume each tile has the uniform power. Power dissipation profile is characterized by the spacial distribution of tile power. Power dissipation of 3D FPGAs is the sum of dynamic power and static power (leakage power). The methodology of modeling the two types of power is introduced as follows.

3.2.1 Dynamic Power

Dynamic power is generated by transition of signals. Signal transition will cause frequently charging and discharging capacitors and this forms the most significant contribution to the dynamic power in 3D FPGAs. This kind of dynamic power can be modeled using the following equation:

$$P = \sum_i C_i V_i^2 D_i f_i \quad (5)$$

where C_i , V_i , D_i and f_i are the total capacitance, swing voltage, signal transition density and operating frequency of source i . Another component of dynamic power is the short-circuit power which is caused by signal switching. According to [21], while short-circuit power in CLB accounts for a higher percentage of its total power, it only contributes less than 10% to the power in the interconnect in an FPGA. In order to capture both factors of dynamic power and avoid complex computation, in our work, we model the dynamic power in two independent parts: Interconnect Power and CLB Power.

Interconnect Power: The calculation of interconnect power takes two steps. First, we compute the power for each routing segment (including horizontal routing segments on each layer and TSVs). A segment connects two terminals (a terminal is an input/output pin of a CLB or a pin of an SB). The equation to calculate the capacitor-charging-based dynamic power of segment i is shown as follows:

$$P_i = C_i V_i^2 D_i f_i \quad (6)$$

where C_i is the sum of input/output capacitance of the two terminals and the distributed capacitance of the segment; D_i is the signal transition density of the net to which the segment belongs and its value is calculated by ACE2.0[10]; f_i is the operating frequency of the system and V_i is simply taken as the supply voltage. Second, we project the power of each segment to the tile because the power of each tile is what really matters. In order to do this, we divide the segment power equally into two parts, and add one part to the tile containing one terminal and add the other part to the tile containing the other terminal. Note that, it is possible that both terminals of a segment reside in one tile. In this case, the whole power of the segment is assigned to this tile.

CLB Power: According to our FPGA model, a CLB contains LUTs, flip-flops, multiplexers, buffers and memory cells. Modeling the dynamic power in a CLB is difficult due to the complex sub-routing within the CLB. In our work, we use a simulation-based method to model the CLB power. First of all, we will calculate the average dynamic power of an active CLB. We use SPICE to simulate each component in a CLB with random input vector pairs at a certain frequency. The simulation gives the dynamic power of each component for all the pairs of input vector. The dynamic power of each component is then taken as the average power for all the input vector pairs of the component, with the assumption that each pair of input vectors has the same probability of occurrence. The dynamic power of a CLB ($Power_{CLB,0}$) is the sum of all the components constituting the CLB. Next, we take $Power_{CLB,0}$ as the input and scale it with the real frequency to get the real dynamic power of an active CLB. Finally, this power is added to the tile which contains the CLB.

3.2.2 Static Power

Static power is caused by the leakage current in FPGAs and can be categorized into two types: gate leakage and source-to-drain leakage. Detailed modeling of the two types of static power is difficult and not accurate at small technology nodes. Therefore, in our work, we take the experiment value from [1] and calculate the leakage power for different temperatures based on an analytical extrapolation function. According to [1], at $85^\circ C$, Typical High-Performance Stratix III has $3W$ static power for 300K logic blocks (similar to the tile in our model). Since the static power is linear to the number of logic blocks, we can calculate the static power for a single logic block, which is around $10\mu W$. This value is then taken as the static power of a 2D-tile at $85^\circ C$ in our 3D FPGA model. The static power of a 3D-tile is $\gamma \times 10\mu W$ where γ is a scaling parameter equal to the ratio of the number of minimum width transistor areas (without considering the area occupied by TSVs) between a 3D-tile and a 2D-tile. Following this, we use the following equation[20] to extrapolate the static power for different temperatures:

$$P_{stat}(T) = P_{stat}(T_0) \times (5.121(\frac{T}{T_0})^2 - 6.013\frac{T}{T_0} + 1.892) \quad (7)$$

Now we get a look-up-table for static power at different temperatures, which is then fed into the simulation framework.

3.3 Thermal Model

The thermal behavior of 3D FPGAs can be characterized with thermal resistance-capacitance (RC) network which resembles the electrical RC network. In order to do this, a 3D FPGA is divided into grids and each grid is represented with a node in the thermal RC network. In this network, the voltage of each node represents the average temperature of the grid while the power of the grid is indicated by the current source connected to the grid. The resistance connecting each adjacent pair of nodes indicates a heat transfer path and the thermal capacitance represents the ability to store heat. In this work, we are mostly interested in steady state thermal behavior. Therefore we will primarily focus on the resistance network. Our resistance network is the same as the one introduced in [23], which is illustrated in Figure 7. The cooling performance of micro-channels is captured by three resistors: R_{cond} , R_{conv} and R_{heat} . R_{cond} and R_{conv} represent the heat transfer between silicon and fluid by conduction and convection respectively. R_{heat} indicates the heat transfer within the fluid. More detailed definition of the resistance network can be found in [23].

After we get the profile of the total power dissipation, we

Figure 7: Thermal model of 3D stacked FPGAs

can calculate the temperature with the following equation:

$$GT = P \quad (8)$$

where G is the thermal conductance matrix determined by the thermal resistance network. During the simulation process, when a new temperature profile is calculated, the static power of each tile is updated according to the new temperature. Then the temperature is recalculated based on the updated total power dissipation. This is illustrated in Figure 6. This loop stops until the temperature is converged.

4. SIMULATION AND DISCUSSION

We will describe our simulation and discuss the results in this section. First, we will introduce the setup of the simulation. Next, we use our analysis framework to perform exploration on 12 benchmarks from the Microelectronics Center of North Carolina (MCNC) benchmark suit and study (1) the different characteristics of air-cooling and micro-channel-based fluidic cooling and (2) the impact of micro-channel density in a 3D FPGA embedded with micro-channel-based fluidic cooling.

4.1 Setup of the Simulation

During our design space exploration, we study how the cooling method and the density of micro-channel influence the electrical properties of a 3D FPGA. By “electrical properties”, we mean the operating frequency (Freq.), Power per CLB (P_{CLB}) and Energy Efficiency (E.E.), where Energy Efficiency is calculated with Equation 4.

We first run the thermo-electrical simulation on fully-vertical-connected 3D FPGAs (100% of the tiles are 3D-tile) with air-cooling. Following this, we apply micro-channel-based fluidic cooling in 3D FPGAs and sweep the micro-channel density. In our work, we change the micro-channel density by increasing the silicon thickness between channels in multiples of the channel width (Figure 5). We define a new parameter, Micro-Channel Pitch (MC_{pitch}), as $MC_{pitch} = \frac{W_{wall}}{W_{ch}}$ to describe the micro-channel density. The vertical bandwidth is maximized for each micro-channel density. Due to the restriction of the size of benchmarks, we range MC_{pitch} from 1 to 11 with a step of 1, which implies that the percentage of 3D-tiles among all tiles ranges from 50% to 92%. For the fully-vertical-connection, only air-cooling is applied due to the lack of space for micro-channels, while for other cases a hybrid cooling scheme (air-cooling and micro-channel-based fluidic cooling) is used. In this work, we assume the temperature limit is $85^\circ C$ [23]. If the maximum temperature exceeds this limit, the design is regarded as thermally infeasible. In this case, we scale down the operating frequency to fix the thermal problem.

Other simulation variables for 3D FPGAs embedded with micro-channel-based fluidic cooling are summarized as follows:

- (1) 3D FPGAs stack up to four functional layers.
- (2) Each 3D-SB placed on the i^{th} layer has four TSVs

Table 1: Electrical and thermal variables employed in the simulation

Variable	Value	Unit	Description
T_{amb}	25	$^{\circ}C$	Ambient temperature
T_{ch}	25	$^{\circ}C$	Inlet temperature of micro-channel
v_f	1	$m\ s^{-1}$	Fluid velocity
k_{ox}	1.4	$W\ m^{-1}K$	Oxide thermal conductivity
k_{si}	149	$W\ m^{-1}K$	Silicon thermal conductivity
k_{cu}	401	$W\ m^{-1}K$	Copper thermal conductivity
k_f	0.607	$W\ m^{-1}K$	Fluid thermal conductivity at $25^{\circ}C$
R_{seg}	101	Ω	Resistance of one routing segment
C_{seg}	22.1	fF	Capacitance of one routing segment
V_{dd}	0.9	V	Supply voltage

connecting to the $(i - 1)^{th}$ layer and $(i + 1)^{th}$ layer, respectively.

- (3) For each TSV, the diameter is $10\mu m$ and the height is $150\mu m$. The resistance and capacitance of a single TSV is $32m\Omega$ and $223fF$.
- (4) According to [25], The dimension of a single micro-channel is set as $50\mu m$ and $100\mu m$ (Width $\times$ Height). This will yield a pressure drop less than $1 \times 10^5 Pa$ and pumping power blow $1W$ according to the model in [23]. Other channel dimensions may have different impact on cooling performance. As illustrated by [23], increasing channel width on one hand reduces the pressure drop hereby leading to smaller thermal resistance, but on the other hand it causes decrease in the coverage of micro-channels hence degrading the heat removal rate. Since the main focus of this paper is the impact of micro-channel density on the performance of 3D FPGAs, we fixed the channel dimension during our simulation although our algorithm works for other channel dimensions as well.
- (5) each CLB is composed of a 4-input LUT, a flip-flop and corresponding memory elements. The width of horizontal routing channel is 30.
- (6) Table 1 lists the rest of the important electrical and thermal variables employed in this work.

Table 4 and Table 5 show the results of some benchmarks on 3D FPGAs before and after scaling down the operating frequency, respectively. Because of the page limits, only seven out of the twelve benchmarks are shown in the tables. The tables illustrate the peak temperature (MaxT) as well as normalized operating frequency (Freq.) and energy efficiency (E.E.) for each benchmark. In Section 4.2 and 4.3, we will discuss the results in detail.

4.2 Fully-Vertical-Connected 3D FPGAs

In fully-vertical-connected 3D FPGAs, all tiles are 3D-tile, thus micro-channels cannot be allocated under this scenario. Only air-cooling is applied during the simulation. Figure 8(a) shows the maximum peak temperature over all benchmarks as well as normalized operating frequency, power per CLB, and energy efficiency for each number of stacked layers. For each benchmark, operating frequency, power per CLB and energy efficiency are normalized to their counterparts when the benchmark is realized on a 2D FPGA. Then the normalized characters are averaged over all the benchmarks to get the results shown in Figure 8(a). Under this scenario, since there is no room to place micro-channels, we can only use air-cooling to remove heat from top of the 3D FPGA. As illustrated by this figure, the peak temperature of some benchmarks exceeds $T_{limit} = 85^{\circ}C$ even when only 2 functional layers are stacked. In addition, the peak temperature increases dramatically by increasing the number of layers in a 3D FPGA. Even though operating frequency and energy efficiency can be improved by stacking more layers in a 3D FPGA, these benefits cannot be realized with a thermally-infeasible design. In order to fix the thermal problem, we scaling down the operating frequency for each

Figure 8: Maximum peak temperature (MaxT), normalized frequency, normalized power per CLB and normalized energy efficiency for different number of layers (a) before fixing thermal problem and (b) after fixing the thermal problem

thermally-violated design and the results are shown in Figure 8(b).

According to the results illustrated in Figure 8(b), due to the thermal restriction, increasing number of layers may lead to the degradation of frequency and energy efficiency. Power per CLB also decreases because of the reduction of operating frequency. This degradation can be reduced by using micro-channel-based fluidic cooling. Micro-channel structure imposes new constraints on 3D FPGA design and we will study the impact of micro-channel density on the electrical properties of 3D FPGAs in Section 4.3.

4.3 3D FPGAs with Micro-channel-based Fluidic Cooling

In this scenario, some 3D-tiles are replaced by 2D-tiles, thus a number of micro-channels can be allocated below 2D-tiles. We will discuss the impact of micro-channel density on the cooling and performance of 3D FPGAs in this section.

4.3.1 Before Fixing Thermal Problem

The trend of operating frequency (Freq.), power per CLB (P_{CLB}) and energy efficiency (E.E.) with increasing micro-channel pitch for different number of stacked layers is shown in Figure 9(a)-(c). Note that increasing micro-channel pitch reduces the micro-channel density thereby reducing the cooling capacity. The approach of normalization of each of the three electrical properties is the same as described in Section 4.2. In each figure, we also show the maximum peak temperature over all benchmarks.

According to Figure 9(a), as we increase the number of layers in a 3D FPGA, the frequency will increase. However, for a certain number of layers, the change of frequency with reducing micro-channel density is not monotonic. In our work, reducing micro-channel density represents the increase in vertical bandwidth. This nonmonotonic relationship between frequency and vertical bandwidth is also discovered in

Figure 9: Maximum peak temperature (MaxT) and normalized (a) operating frequency, (b) power per CLB (P_{CLB}), (c) energy efficiency for different number of layers and micro-channel density before fixing the thermal problem

Figure 10: Maximum peak temperature (MaxT) and normalized (a) operating frequency, (b) power per CLB (P_{CLB}), (c) energy efficiency for different number of layers and micro-channel density after fixing the thermal problem

[24]. This phenomenon can be caused by the following two factors:

- (1) Although increasing vertical bandwidth will improve the latency in 3D nets, it can also cause congestion in 2D routing channels around the 3D-SB which may lead to the increase in delay of other nets.
- (2) The distribution of routing resources changes when we alter the micro-channel density, which forces the router to find completely new paths to connect CLBs. Since each routing channel has tracks of different lengths, these paths may have very significant variations in the RC parameters which are essential in calculating delay.

Since both power density and energy efficiency depend on the operating frequency, they also exhibit nonmonotonic behavior as shown in Figure 9(b)-(c).

Another interesting point that can be made is about power per CLB (P_{CLB}). According to Figure 9(b), for a certain number of layers, power per CLB has an increasing tendency when reducing the micro-channel density. This can be explained as follows: reducing micro-channel density leads to lower mass flow rate of coolant which will reduce the cooling capacity of micro-channel heat sink and increase the on-chip temperature; increased on-chip temperature will in turn cause the rise of leakage power, thus increasing the power dissipation for each CLB. On the other hand, for a fixed micro-channel density, when another functional layer is added, the power dissipation per CLB will increase. This behavior is determined by three factors: (1) operating frequency may increase with the number of layers which leads to the increase of dynamic power per CLB; (2) increasing the number of stacked layers enables us to realize a certain benchmark using smaller number of CLBs; (3) increasing the number of layers causes the rise of on-chip temperature thereby increasing the leakage power per CLB due to the positive feedback between temperature and leakage power.

Since increasing power per CLB may imply larger power dissipation and energy efficiency is inversely related to power, there is a decreasing tendency in energy efficiency by increasing micro-channel pitch. As illustrated in Figure 9(c), this tendency becomes more obvious when the number of layers increases. By increasing the number of layers for a fixed micro-channel density, however, energy efficiency will increase since operating frequency is dominated.

4.3.2 After Fixing Thermal Problem

According to the results shown in Figure 9, although micro-channel-based fluidic cooling can significantly reduce the on-chip temperature. if micro-channel density is too small, thermal violations can still occur. For these cases, we scale down the operating frequency to fix the thermal problem and the results are shown in Figure 10(a)-(c). The figures illustrate that the peak temperatures of all designs are below T_{limit} and the curves of electrical properties are pulled down at the high-micro-channel-pitch end due to the reduction of operating frequency. According to Figure 10(a), the drop of frequency becomes sharper and occurs earlier when more layers are stacked in 3D FPGAs. This is because increasing the number of layers intensifies the thermal problem of 3D FPGAs and frequency should be scaled down to a much lower level in order to fix the thermal problem. The degradation of frequency meanwhile causes the reduction in power dissipation per CLB and energy efficiency. According to the figures, we can get the following guidelines for designing 3D FPGAs embedded with micro-channel-based fluidic cooling: (1) the performance of 3D FPGAs with higher vertical bandwidth will be “locked” due to the thermal restriction; (2) there is an “unlocked window” close to the lower-vertical-bandwidth end where we can find the optimal design of a 3D FPGA; (3) the size of unlocked window shrinks as increasing

Table 2: Comparison to Best Cooling Design

	2 Layers	3 Layers	4 Layers
Freq.	15.0%	20.2%	4.8%
E.E.	15.1%	23.8%	6.0%

Table 3: Comparison to Best Bandwidth Design

	2 Layers	3 Layers	4 Layers
Freq.	24.3%	47.3%	80.3%
E.E.	35.5%	74.0%	124.0%

the number of layers; (4) the architecture of optimal design is different for different number of layers.

Our analysis framework also identifies the optimal physical architecture of 3D FPGAs with different number of stacked layers for each benchmark. In order to evaluate the optimal designs, we compare their operating frequency and energy efficiency against that of the designs with largest micro-channel density (Best Cooling) and smallest micro-channel density (Best Bandwidth), respectively. The improvement is averaged over all the benchmarks and the results are shown in Table 2 and Table 3. According to the tables, the improvement with respect to the Best Bandwidth Design is generally larger than that with respect to the Best Cooling Design. This is because that the frequency of the Best Bandwidth Design has already been scaled down to fix the thermal problem, thus leading to the degradation of performance and energy efficiency. It should be noted that optimal design varies for different benchmarks and different number of layers. The optimal design of some benchmarks are shown in bold type in Table 5.

5. CONCLUSION

In this work, we studied the thermo-electrical interaction in 3D FPGAs. Our results demonstrate that micro-channel based fluidic cooling is necessary to realize the true potential of 3D integration technology. We proposed an analysis framework to study the impact of micro-channel density on the performance and identify the optimal 3D FPGA design for each benchmark. According to the simulation results, we proposed guidelines for designing the physical architecture of 3D FPGAs embedded with micro-channel-based fluidic cooling. Comparison with Best Cooling Design and Best Bandwidth Design shows that the optimal design identified by our framework can improve operating frequency and energy efficiency up to 80.3% and 124.0% respectively.

6. ACKNOWLEDGEMENTS

The authors acknowledge that this work has been funded by NSF grant CCF1302375 and DARPA ICECOOL.

7. REFERENCES

- [1] Peter Cheung. 2008. Modern FPGA architectures. (Jun 2008). Retrieved August 30, 2015. from <http://www.ee.ic.ac.uk/>.
- [2] Virtex-4 family overview. (Aug 30 2010) Retrieved September 5, 2015 from <http://www.xilinx.com>.
- [3] C. Ababei, et al. Exploring potential benefits of 3D FPGA integration. In *Field programmable logic and application*, pages 874–880, 2004.
- [4] J. H. Anderson and F. N. Najm. Active leakage power optimization for FPGAs. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 25(3):423–437, 2006.
- [5] V. Betz, et al. *Architecture and CAD for deep-submicron FPGAs*, volume 497. 2012.
- [6] A. Gayasen, et al. Designing a 3-D FPGA: switch box architecture and thermal issues. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 16(7):882–893, 2008.
- [7] D. H. Kim, et al. A study of Through-Silicon-Via impact on the 3D stacked IC layout. In *ICCAD*, pages 674–680, Nov 2009.
- [8] J.-M. Koo, et al. Integrated microchannel cooling for three-dimensional electronic circuit architectures. *Journal of heat transfer*, 127(11):49–58, Jan 2005.
- [9] I. Kuon, et al. Fpga architecture: Survey and challenges. *Foundations and Trends in Electronic Design Automation*, 2(2):135–253, 2008.
- [10] J. Lamoureux and S. J. Wilton. Activity estimation for field-programmable gate arrays. In *Field Programmable Logic and Applications, 2006. FPL’06. International Conference on*, page 1.
- [11] Y. Y. Liao, et al. Nonvolatile 3D-FPGA with monolithically stacked RRAM-based configuration memory. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pages 406–408, 2012.
- [12] M. Lin, et al. Performance benefits of monolithically stacked 3-D FPGA. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 26(2):216–229, 2007.
- [13] S. Ndao, et al. Multi-objective thermal design optimization and comparative analysis of electronics cooling technologies. *International Journal of Heat and Mass Transfer*, 52(19):4317–4326, 2009.
- [14] V. Pangracious, et al. Designing a 3D tree-based FPGA: Optimization of butterfly programmable interconnect topology using 3D technology. In *3D Systems Integration Conference (3DIC), 2013 IEEE International*, pages 1–8, 2013.
- [15] S. A. Razavi, et al. A tileable switch module architecture for homogeneous 3D FPGAs. In *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, pages 1–4, 2009.
- [16] J. Ryu, et al. Numerical optimization of the thermal performance of a microchannel heat sink. *International Journal of Heat and Mass Transfer*, 45(13):2823–2827, 2002.
- [17] M. Sabry, et al. GreenCool: an energy-efficient liquid cooling design technique for 3-D MPSoCs Via channel width modulation. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, 32(4):524–537, 2013.
- [18] T. E. Sarvey, et al. Embedded cooling technologies for densely integrated electronic systems. In *Custom Integrated Circuits Conference (CICC), 2015 IEEE*, pages 1–8, 2015.
- [19] I. Savidis, et al. Electrical modeling and characterization of through-silicon vias (TSVs) for 3-D integrated circuits. *Microelectronics Journal*, 41(1):9–16, 2010.
- [20] C. Serafy, et al. Continued frequency scaling in 3D ICs through micro-fluidic cooling. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014 IEEE Intersociety Conference on*, pages 79–85, 2014.
- [21] L. Shang, et al. Dynamic power consumption in Virtex-II FPGA family. In *Proceedings of the 2002 ACM/SIGDA tenth international symposium on Field-programmable gate arrays*, pages 157–164, 2002.
- [22] B. Shi, et al. Hybrid 3D-IC cooling system using micro-fluidic cooling and thermal TSVs. In *ISVLSI*, pages 33–38, August 2012.
- [23] B. Shi, et al. Non-uniform micro-channel design for stacked 3D-ICs. In *DAC*, pages 658–663, June 2011.
- [24] K. Siozios, et al. Architecture-level exploration of alternative interconnection schemes targeting 3d fpgas: A software-supported methodology. *International Journal of Reconfigurable Computing*, 2008, 2009.
- [25] A. Sridhar, et al. 3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling. In *Proceedings of the International Conference on Computer-Aided Design*, pages 463–470, 2010.
- [26] T. Tuan, et al. A 90nm low-power FPGA for battery-powered applications. In *Proceedings of the 2006 ACM/SIGDA 14th international symposium on Field programmable gate arrays*, pages 3–11, 2006.
- [27] D. Tuckerman and R. Pease. High-performance heat sinking for VLSI. *IEEE Electron Device Lett.*, 2(5):126–129, May 1981.
- [28] X. Wei and Y. Joshi. Stacked microchannel heat sinks for liquid cooling of microelectronic components. *Journal of Electronic Packaging*, 126(1):60–66, 2004.

Table 4: Results of some benchmarks on 3D FPGAs before scaling down the operating frequency (FVC = Fully-Vertical-Connection Scenario, MaxT = Peak Temperature, Freq. = Frequency, E.E. = Energy Efficiency, MC_{pitch} = Micro-channel Pitch. MaxT has the unit of $^{\circ}C$, Freq. and E.E. are normalized to their counterparts on 2D FPGAs. Thermally infeasible designs are shown in bold type.)

			MC _{pitch}											
			FVC	1	2	3	4	5	6	7	8	9	10	11
alu4	2 Layers	maxT	89.45	43.27	47.06	50.90	59.55	59.80	61.94	61.00	67.13	74.53	71.91	71.43
		Freq.	1.72	1.55	1.50	1.52	1.90	1.62	1.63	1.44	1.70	1.97	1.74	1.63
		E.E.	1.72	1.74	1.64	1.70	2.22	1.73	1.79	1.51	1.87	2.22	1.89	1.73
	3 Layers	maxT	164.42	42.91	54.93	55.81	63.47	68.34	69.84	75.99	76.30	75.38	85.59	86.67
		Freq.	2.34	1.81	2.57	2.04	2.30	2.31	2.15	2.26	2.15	1.91	2.15	2.18
		E.E.	2.16	2.18	3.35	2.47	2.89	2.86	2.59	2.72	2.60	2.22	2.46	2.57
4 Layers	maxT	237.27	47.44	54.41	60.53	66.48	71.39	77.74	81.41	84.33	98.80	96.58	100.64	
	Freq.	2.74	2.79	2.73	2.86	2.81	2.71	2.76	2.74	2.59	3.05	2.74	2.74	
	E.E.	2.79	3.99	3.82	4.19	4.01	3.78	3.76	3.79	3.44	4.12	3.62	3.58	
bigkey	2 Layers	maxT	121.80	56.03	66.23	65.94	69.58	82.58	82.89	89.92	92.11	95.76	98.32	98.70
		Freq.	3.28	3.49	3.68	2.98	2.93	3.51	3.26	3.46	3.41	3.45	3.46	3.33
		E.E.	5.70	7.15	7.49	5.82	5.69	6.86	6.30	6.69	6.54	6.58	6.60	6.24
	3 Layers	maxT	188.79	53.44	61.70	69.42	84.60	75.45	91.49	94.67	89.17	117.92	113.00	127.89
		Freq.	3.14	3.10	3.05	3.05	3.68	2.62	3.26	3.14	2.62	3.73	3.31	3.73
		E.E.	3.83	5.07	4.89	4.77	5.95	3.92	5.03	4.78	3.84	5.64	4.89	5.52
4 Layers	maxT	293.59	57.54	63.19	70.96	88.38	94.29	103.15	108.02	123.98	128.61	131.79	138.32	
	Freq.	4.45	4.53	4.02	3.85	4.64	4.43	4.49	4.24	4.78	4.68	4.45	4.45	
	E.E.	7.05	9.49	8.17	7.65	9.30	8.76	8.77	8.06	9.09	8.86	8.25	8.16	
des	2 Layers	maxT	83.45	44.51	45.30	52.06	56.60	59.71	65.39	59.23	66.77	62.91	69.51	72.23
		Freq.	3.22	3.21	2.50	3.03	3.15	3.21	3.62	2.64	3.26	2.64	3.21	3.31
		E.E.	4.79	5.43	3.92	5.03	5.13	5.24	6.11	4.06	5.27	3.97	5.12	5.26
	3 Layers	maxT	163.47	48.79	54.81	60.12	66.65	72.93	76.17	78.47	81.77	86.21	91.77	86.75
		Freq.	4.28	4.78	4.63	4.36	4.61	4.65	4.60	4.32	4.23	4.34	4.56	3.85
		E.E.	5.75	9.49	9.11	8.27	8.87	8.71	8.65	7.88	7.55	7.76	8.13	6.59
4 Layers	maxT	254.74	50.18	61.36	66.26	72.94	83.57	90.47	87.30	92.93	102.13	107.66	113.49	
	Freq.	5.67	5.65	6.32	5.76	5.69	6.22	6.09	5.21	5.29	5.62	5.67	5.67	
	E.E.	9.30	12.75	14.41	12.82	8.87	13.76	13.01	10.87	11.01	11.62	11.62	11.43	
diffee	2 Layers	maxT	83.10	44.04	47.54	51.50	57.01	59.73	60.82	64.09	67.20	72.41	72.03	68.64
		Freq.	1.72	1.72	1.58	1.63	1.68	1.72	1.64	1.73	1.73	1.91	1.81	1.59
		E.E.	1.98	2.24	1.99	2.11	2.11	2.20	2.06	2.19	2.14	2.41	2.25	1.93
	3 Layers	maxT	154.55	44.30	53.26	54.22	61.72	62.16	73.76	77.32	74.46	86.30	87.74	87.77
		Freq.	2.46	2.08	2.42	1.99	2.21	1.95	2.45	2.37	2.04	2.42	2.36	2.27
		E.E.	2.65	2.97	3.54	2.75	3.09	2.63	3.44	3.23	2.69	3.26	3.17	3.03
4 Layers	maxT	208.64	44.27	51.52	56.12	61.43	69.00	71.20	81.10	87.93	83.42	98.15	98.91	
	Freq.	2.71	2.44	2.47	2.43	2.41	2.56	2.44	2.68	2.92	2.39	2.91	2.71	
	E.E.	2.71	3.53	3.45	3.41	3.30	3.54	3.30	3.64	4.04	3.10	3.90	3.52	
dsip	2 Layers	maxT	107.80	48.12	54.54	63.37	70.77	68.22	74.19	78.43	80.36	80.43	88.56	87.33
		Freq.	2.14	2.05	2.09	2.37	2.51	2.08	2.24	2.29	2.23	2.12	2.38	2.24
		E.E.	3.17	3.55	3.62	4.12	4.39	3.49	3.79	3.84	3.70	3.46	3.91	3.64
	3 Layers	maxT	187.01	47.86	53.24	57.23	75.21	67.87	78.02	85.30	94.11	97.61	101.70	103.37
		Freq.	2.76	2.47	2.33	2.17	3.02	2.19	2.54	2.68	2.88	2.83	2.87	2.76
		E.E.	4.03	4.77	4.42	4.02	5.82	3.95	4.69	4.94	5.26	5.15	5.23	4.93
4 Layers	maxT	290.81	55.59	67.26	73.29	83.42	85.09	100.05	97.68	114.66	112.07	118.76	124.38	
	Freq.	3.48	3.85	4.12	3.83	3.88	3.41	3.88	3.38	3.88	3.48	3.48	3.48	
	E.E.	5.71	8.59	9.19	8.34	8.36	7.07	8.11	6.89	7.90	6.94	6.84	6.78	
ex1010	2 Layers	maxT	78.46	45.56	53.45	54.04	58.18	61.40	62.91	64.69	67.81	65.74	68.92	71.77
		Freq.	1.93	1.57	1.95	1.66	1.73	1.78	1.75	1.79	1.90	1.66	1.79	1.85
		E.E.	2.30	1.97	2.62	2.12	2.18	2.26	2.19	2.27	2.45	2.03	2.23	2.30
	3 Layers	maxT	144.43	51.86	58.72	65.91	70.12	75.61	81.78	81.81	82.42	87.53	95.80	103.08
		Freq.	2.71	2.68	2.58	2.69	2.54	2.62	2.68	2.46	2.29	2.46	2.68	2.94
		E.E.	3.03	4.14	3.86	4.07	3.71	3.82	3.87	3.47	3.09	3.43	3.73	4.18
4 Layers	maxT	202.35	50.92	58.65	66.59	72.77	76.34	84.75	90.46	91.81	99.10	107.15	115.23	
	Freq.	2.91	2.86	2.81	2.88	2.82	2.65	2.82	2.87	2.61	2.80	2.96	3.10	
	E.E.	3.08	4.54	4.36	4.49	4.25	3.88	4.11	4.17	3.62	3.97	4.17	4.34	
frisc	2 Layers	maxT	77.73	44.05	49.92	52.92	55.33	54.71	60.62	60.17	68.23	62.64	68.36	69.70
		Freq.	2.20	1.88	2.03	1.98	1.91	1.64	1.99	1.81	2.30	1.78	2.10	2.12
		E.E.	2.74	2.52	2.75	2.66	2.52	2.03	2.61	2.29	3.12	2.21	2.75	2.77
	3 Layers	maxT	119.51	46.32	49.95	56.15	61.73	65.31	67.72	69.22	70.69	80.23	81.30	85.67
		Freq.	2.48	2.50	2.25	2.40	2.47	2.46	2.36	2.24	2.15	2.59	2.49	2.63
		E.E.	2.64	3.59	3.13	3.35	3.42	3.38	3.17	2.93	2.76	3.47	3.26	3.49
4 Layers	maxT	191.26	46.60	50.42	59.96	68.63	70.68	73.30	80.59	78.44	89.19	86.14	99.56	
	Freq.	2.91	2.93	2.54	3.00	3.26	2.94	2.79	2.96	2.53	2.95	2.57	3.11	
	E.E.	2.94	4.64	3.86	4.70	5.10	4.43	4.11	4.39	3.53	4.24	3.53	4.43	

Table 5: Results of some benchmarks on 3D FPGAs after scaling down the operating frequency (FVC = Fully-Vertical-Connection Scenario, MaxT = Peak Temperature, Freq. = Frequency, E.E. = Energy Efficiency, MC_{pitch} = Micro-channel Pitch. MaxT has the unit of $^{\circ}C$, Freq. and E.E. are normalized to their counterparts on 2D FPGAs. Optimal designs are shown in bold type.)

			MC_{pitch}											
			FVC	1	2	3	4	5	6	7	8	9	10	11
alu4	2 Layers	maxT	84.38	43.27	47.06	50.90	59.55	59.80	61.94	61.00	67.13	74.53	71.91	71.43
		Freq.	1.55	1.55	1.50	1.52	1.90	1.62	1.63	1.44	1.70	1.97	1.74	1.63
		E.E.	1.52	1.74	1.64	1.70	2.22	1.73	1.79	1.51	1.87	2.22	1.89	1.73
	3 Layers	maxT	84.33	42.91	54.93	55.81	63.47	68.34	69.84	75.99	76.30	75.38	84.49	84.92
		Freq.	0.69	1.81	2.57	2.04	2.30	2.31	2.15	2.26	2.15	1.91	2.10	2.10
		E.E.	0.44	2.18	3.35	2.47	2.89	2.86	2.59	2.72	2.60	2.22	2.39	2.46
	4 Layers	maxT	83.80	47.44	54.41	60.53	66.48	71.39	77.74	81.41	84.33	84.37	84.29	84.43
		Freq.	0.17	2.79	2.73	2.86	2.81	2.71	2.76	2.74	2.59	2.34	2.17	2.03
		E.E.	0.04	3.99	3.82	4.19	4.01	3.78	3.76	3.79	3.44	3.03	2.75	2.52
bigkey	2 Layers	maxT	83.96	56.03	66.23	65.94	69.58	82.58	82.89	84.12	84.62	84.54	84.58	84.56
		Freq.	1.77	3.49	3.68	2.98	2.93	3.51	3.26	3.10	2.97	2.81	2.71	2.58
		E.E.	2.74	7.15	7.49	5.82	5.69	6.86	6.30	5.88	5.58	5.19	4.96	4.64
	3 Layers	maxT	84.32	53.44	61.70	69.42	84.60	75.45	84.29	84.82	84.42	84.49	84.62	84.59
		Freq.	0.71	3.10	3.05	3.05	3.68	2.62	2.84	2.61	2.39	2.19	2.06	1.94
		E.E.	0.54	5.07	4.89	4.77	5.95	3.92	4.29	3.86	3.43	3.05	2.81	2.57
	4 Layers	maxT	83.83	57.54	63.19	70.96	84.55	84.90	84.76	84.51	84.60	84.75	84.47	84.38
		Freq.	0.19	4.53	4.02	3.85	4.32	3.74	3.29	2.87	2.65	2.45	2.23	2.06
		E.E.	0.06	9.49	8.17	7.65	8.59	7.24	6.16	5.15	4.63	4.21	3.70	3.35
des	2 Layers	maxT	83.45	44.51	45.30	52.06	56.60	59.71	65.39	59.23	66.77	62.91	69.51	72.23
		Freq.	3.22	3.21	2.50	3.03	3.15	3.21	3.62	2.64	3.26	2.64	3.21	3.31
		E.E.	4.79	5.43	3.92	5.03	5.13	5.24	6.11	4.06	5.27	3.97	5.12	5.26
	3 Layers	maxT	83.87	48.79	54.81	60.12	66.65	72.93	76.17	78.47	81.77	84.47	84.36	84.72
		Freq.	1.25	4.78	4.63	4.36	4.61	4.65	4.60	4.32	4.23	4.19	3.94	3.69
		E.E.	1.15	9.49	9.11	8.27	8.87	8.71	8.65	7.88	7.55	7.43	6.83	6.17
	4 Layers	maxT	83.82	50.18	61.36	66.26	72.94	83.57	84.46	84.54	84.39	84.68	84.58	84.46
		Freq.	0.31	5.65	6.32	5.76	5.69	6.22	5.44	4.94	4.50	4.13	3.81	3.50
		E.E.	0.11	12.75	14.41	12.82	12.56	13.76	11.42	10.21	9.10	8.08	7.29	6.48
diffeq	2 Layers	maxT	83.10	44.04	47.54	51.50	57.01	59.73	60.82	64.09	67.20	72.41	72.03	68.64
		Freq.	1.72	1.72	1.58	1.63	1.68	1.72	1.64	1.73	1.73	1.91	1.81	1.59
		E.E.	1.98	2.24	1.99	2.11	2.11	2.20	2.06	2.19	2.14	2.41	2.25	1.93
	3 Layers	maxT	84.23	44.30	53.26	54.22	61.72	62.16	73.76	77.32	74.46	84.82	84.72	84.60
		Freq.	0.84	2.08	2.42	1.99	2.21	1.95	2.45	2.37	2.04	2.34	2.22	2.13
		E.E.	0.68	2.97	3.54	2.75	3.09	2.63	3.44	3.23	2.69	3.14	2.95	2.80
	4 Layers	maxT	83.89	44.27	51.52	56.12	61.43	69.00	71.20	81.10	84.88	83.42	84.40	84.41
		Freq.	0.31	2.44	2.47	2.43	2.41	2.56	2.44	2.68	2.75	2.39	2.25	2.06
		E.E.	0.11	3.53	3.45	3.41	3.30	3.54	3.30	3.64	3.77	3.10	2.86	2.53
dsip	2 Layers	maxT	84.32	48.12	54.54	63.37	70.77	68.22	74.19	78.43	80.36	80.43	84.62	84.88
		Freq.	1.41	2.05	2.09	2.37	2.51	2.08	2.24	2.29	2.23	2.12	2.20	2.14
		E.E.	1.93	3.55	3.62	4.12	4.39	3.49	3.79	3.84	3.70	3.46	3.59	3.45
	3 Layers	maxT	84.34	47.86	53.24	57.23	75.21	67.87	78.02	84.90	84.56	84.98	85.00	84.96
		Freq.	0.64	2.47	2.33	2.17	3.02	2.19	2.54	2.66	2.41	2.25	2.14	2.00
		E.E.	0.58	4.77	4.42	4.02	5.82	3.95	4.69	4.89	4.28	3.94	3.69	3.37
	4 Layers	maxT	84.07	55.59	67.26	73.29	83.42	84.78	84.88	84.97	84.81	84.96	84.62	84.74
		Freq.	0.16	3.85	4.12	3.83	3.88	3.39	3.00	2.70	2.43	2.25	2.05	1.91
		E.E.	0.05	8.59	9.19	8.34	8.36	7.02	6.07	5.33	4.64	4.21	3.71	3.39
ex1010	2 Layers	maxT	78.46	45.56	53.45	54.04	58.18	61.40	62.91	64.69	67.81	65.74	68.92	71.77
		Freq.	1.93	1.57	1.95	1.66	1.73	1.78	1.75	1.79	1.90	1.66	1.79	1.85
		E.E.	2.30	1.97	2.62	2.12	2.18	2.26	2.19	2.27	2.45	2.03	2.23	2.30
	3 Layers	maxT	84.03	51.86	58.72	65.91	70.12	75.61	81.78	81.81	82.42	84.36	84.31	84.94
		Freq.	0.96	2.68	2.58	2.69	2.54	2.62	2.68	2.46	2.29	2.30	2.11	2.07
		E.E.	0.77	4.14	3.86	4.07	3.71	3.82	3.87	3.47	3.09	3.15	2.77	2.72
	4 Layers	maxT	83.96	50.92	58.65	66.59	72.77	76.34	84.75	84.99	84.87	84.97	84.59	84.53
		Freq.	0.26	2.86	2.81	2.88	2.82	2.65	2.82	2.56	2.26	2.11	1.93	1.78
		E.E.	0.07	4.54	4.36	4.49	4.25	3.88	4.11	3.61	3.02	2.78	2.43	2.16
frisc	2 Layers	maxT	77.73	44.05	49.92	52.92	55.33	54.71	60.62	60.17	68.23	62.64	68.36	69.70
		Freq.	2.20	1.88	2.03	1.98	1.91	1.64	1.99	1.81	2.30	1.78	2.10	2.12
		E.E.	2.74	2.52	2.75	2.66	2.52	2.03	2.61	2.29	3.12	2.21	2.75	2.77
	3 Layers	maxT	83.97	46.32	49.95	56.15	61.73	65.31	67.72	69.22	70.69	80.23	81.30	84.99
		Freq.	1.19	2.50	2.25	2.40	2.47	2.46	2.36	2.24	2.15	2.59	2.49	2.59
		E.E.	0.97	3.59	3.13	3.35	3.42	3.38	3.17	2.93	2.76	3.47	3.26	3.42
	4 Layers	maxT	83.97	46.60	50.42	59.96	68.63	70.68	73.30	80.59	78.44	84.49	84.64	84.50
		Freq.	0.30	2.93	2.54	3.00	3.26	2.94	2.79	2.96	2.53	2.67	2.48	2.30
		E.E.	0.09	4.64	3.86	4.70	5.10	4.43	4.11	4.39	3.53	3.74	3.37	3.02