Microelectronic Engineering 91 (2012) 24–32

Contents lists available at SciVerse ScienceDirect

Microelectronic Engineering

journal homepage: www.elsevier.com/locate/mee

A thermal–mechanical coupled finite element model with experimental temperature verification for vertically stacked FPGAs

Chunbo (Sam) Zhang [a], Ramachandra Kallam [b], Andrew Deceuster [a], Aravind Dasu [b], Leijun Li [a,*]

[a] Department of Mechanical & Aerospace Engineering, Utah State University, Logan, UT 84322-4130, USA
[b] Department of Electrical & Computer Engineering, Utah State University, Logan, UT 84322-4120, USA

ARTICLE INFO

Article history:
Received 27 February 2009
Accepted 11 November 2011
Available online 23 November 2011

Keywords:
Thermo-mechanical modeling
Finite element
Three-dimensional fields
ICs
FPGAs

ABSTRACT

Back end of line 3D integration of dies is a promising technology that can allow for considerable boost in inter-chip communication and reduction in form factor of a package. This can result in, however, high die temperatures, particularly for multi-tier FPGAs, due to the high density of power dissipating circuits. Therefore, to design thermally aware multi-tier FPGAs, there is a need to first understand the relationship among circuit architectures, toggle rates, layout, clock frequency, I/O behavior, power, temperature, thermal stress, and thickness of inter-die substrates. This study investigated the relationship among these parameters as tested on a Spartan 3E-250K FPGA. The power-temperature simulations have been conducted on a model built in ANSYS finite element package. The model predictions have been verified by thermocouple-measured case temperatures. The verified model has been extended to predict the temperature and stress distributions for a 2-tier, stacked package.

1. Introduction

3D integration of devices, CPUs, DRAM, FLASH, FPGAs, etc., is a promising technology that can allow for considerable boost in inter-chip communication and reduction in form factor of a package. A surge in publications over the past few years in this area indicates the growing interest in addressing the challenges of 3D integration [1–4]. While 3D packaging itself is not new, it has been restricted mostly to DRAM and FLASH memory products through high precision wire bonding. The sole advantage of wire bonding based vertical integration is reduction in form factor. A new method to carry out vertical integration is commonly referred to as Through Silicon Via (TSV) [5–7]. The benefits of stacking dies with TSVs include: (1) increase in inter-chip bandwidth, (2) reduction in I/O power, and (3) reduction in form factor. A well-acknowledged problem with TSV based multi-tier stacks is intense heat buildup due to higher power densities. This problem is in fact expected to be severe for vertically stacked logic type of dies (CPU, FPGA, etc.) [8].

The thermal issues have become one of the more critical challenges of 3D integrated circuit design. In answer to this problem, different thermal models have recently been developed. The resistive thermal network and electrical-structure-based models are efficient, but generally have a lower accuracy [5,9,1,2,10,11]. Another type of model used has been the finite element method (FEM) or finite difference method (FDM), which was more accurate in directly solving the heat equation, but more time-consuming and CPU-intensive [12,13,6].

Cong et al. proposed an efficient 3D multilevel routing approach that included a novel TSV planning algorithm. The proposed approach featured an adaptive, lumped resistive thermal model and a two-step multilevel TSV planning scheme. Compared to a post processing approach for dummy TSV insertion, their approach used 80% fewer TSVs to achieve the same required temperature [5]. Akturk et al. developed a thermal resistor networks model to simulate the stacked Pentium III processor chips. The number of stacks was limited to 6, due to the increase in maximum processor temperature from the number of layers [10,11]. Puttaswamy and Loh estimated the temperatures of a planar IC based on the Alpha 21364 processor, in addition to a 2-die and 4-die 3D implementation of the same IC. Compared to the planar IC, the 2-die implementation increased the maximum temperature by 17 and 33 K, respectively [14]. Im and Banerjee modeled 3D ICs that were vertically stacked and adhered together with polyimide. Their predicted maximum die temperature was in the range of 380–400 °C [1]. Bentz et al. developed solid 3D stress–strain models in FEMLAB™ 3.1 using the Structural Mechanics Module. They did simulations of 3D IC structures based upon benzocyclobutene (BCB) bonded wafers. Their results of the von Mises stresses supported a goal of using thin BCB [7].

Researchers in the FPGA community had been estimating the die temperatures and power consumptions by the chips. Jones et al. [15] presented a dynamic thermal management system that

* Corresponding author. Tel.: +1 435 797 8184; fax: +1 435 797 2417.
E-mail address: leijun.li@usu.edu (L. Li).

0167-9317/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.mee.2011.11.011

continuously monitored the temperature of the FPGA and shut it down if it reached a particular threshold value. Shang et al. [16] estimated the power consumption of various device resources of an FPGA and gave the distribution of power dissipation for these resources. Power (P) was estimated using the formula, $P = \Sigma fCV^2$, where f was the operating frequency, C was the capacitance, and V was the operating voltage, for a given resource. To do so, resource utilization, switching activity and effective capacitance for each resource were calculated. Resource utilization and switching activity were obtained using Xilinx™ tools and Modelsim. Effective capacitance was calculated in two ways: measurement and simulation. For the measurement, two designs were created; one was the reference design and the other was the reference design with the target resource. The powers of these two designs were subtracted to obtain the power of the target resource and hence the effective capacitance. The simulation was done by generating a netlist for each design in cadence and simulating it in Hspice™. These results were compared with the measurements. The simulation was used to find out the capacitance of the resources whose capacitance could not be isolated for measurement. Results showed that interconnects dissipate the majority of the power (60%), followed by logic (16%) and clocking (14%). The authors stated that the power consumed by the IOBS (Input Output Buffers) was negligible, which may not be the case since IOBs have the highest effective load capacitance compared to other resources on the FPGA.

Lopez et al. [17–19] used a technique proposed by Quenot et al. [20] by implementing ring-oscillators on FPGAs by using the configuration port and read-back capabilities of an FPGA. The output frequency of each ring-oscillator was a function of temperature. The associated counter stored a value that was proportional to the frequency output of the ring-oscillator, and therefore, to the die temperature. This counter was read-back using the configuration port to calculate the actual die temperature.

The thermal sensor (ring-oscillator) was calibrated by placing a thermocouple in the center of the package and measuring the chip temperature while the FPGA board was kept in the temperature-controlled oven. It was not shown, however, how the thermal sensors that measured the die temperature were calibrated against the case temperature measured by the thermocouple. It is not certain that this method was accurate enough to model temperatures and thermal stress of multi-stack 3D FPGAs.

To accurately predict the die temperature and thermal stress, an experiment-and-simulation combined methodology bas been proposed. First, a multi-physics finite element (FE) model, using ANSYS™, has been developed for an FPGA. This model was verified by the experimentally measured temperature data, which were obtained by embedding a thermocouple into the epoxy mold. A parametric study on the relationship between the critical factors of toggle rate, clock frequency, slices utilization, output buffers, and silicon substrate thickness, and die temperature was carried out. The verified FE model for single-FPGA has been extended to predict the thermal and structural behavior of a 2-tier stack of Spartan 3E 250 K FPGAs enclosed in a similar package (TQ144).

2. Experimental procedure

2.1. Circuit designs

Measuring the power consumed by FPGAs was a challenge for two reasons: (1) features of a circuit mapped onto the device, such as the types of resources used, slices, flip flops, BRAMs, multipliers, I/O buffers etc., number of resources used (RU), clock frequency, layout of the circuit, toggle rate (TR) and output pins being driven, all affected the dynamic power consumptions. (2) The vendor supplied power estimation tool (Xpower™ from Xilinx) was inaccurate and unreliable for accurate modeling of thermal behavior. Therefore, two basic building blocks were designed with an approach similar to Jones et al. [15]. Based on these two basic building blocks, various circuits were designed to identify the individual and grouped power contributions of all the above-mentioned parameters. Building blocks BB1 and BB2 are shown in Fig. 1.

BB1 consisted of four input generators. Each input generator consisted of two inverters in a feed-back loop separated by D flip-flops (FF) as shown in Fig. 2. The FFs were all reset to 0 when the chip was powered on. The module generated a simple clock-like signal whose duty cycle could be increased by integer multiples, by increasing the number of FFs. This control enabled the monitoring of power's dependence on toggle rate. It should be noted that such a signal could also be generated by an on-chip digital clock manager (DCM). The DCM was not used because DCMs drive a large number of wires on the clock tree and would dominate power consumption. DCMs were also vendor and family specific.

BB2 consisted of a set of four NAND gates and four FFs, arranged in four rows. The NAND gate was chosen for two reasons. First, when all the inputs were high, its output was low, and when all inputs were low, its output was high. Second, when mapped onto a 4:1 look up table (LUT) on the FPGA, all input lines of the LUT were toggled. The outputs of each NAND gate was registered to prevent the vendor's EDA (electronic design automation) tool from altering the logic, and the NAND gate was the only other generic device primitive type in a slice. The outputs of the four input generators

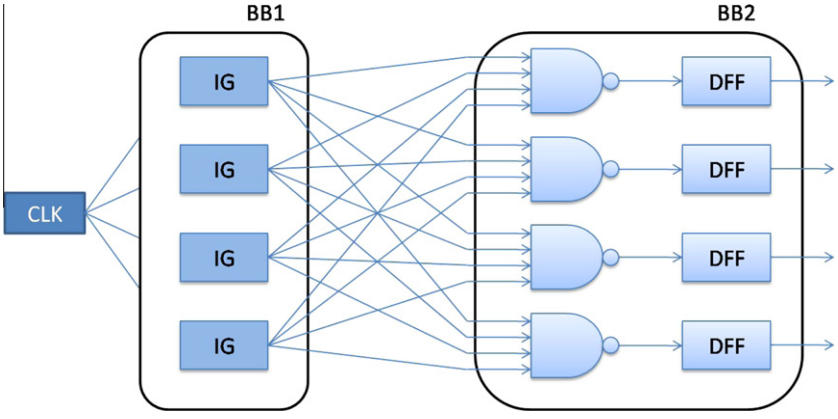

Fig. 1. Building blocks, BB1 and BB2.