Improving Performance in Near-Threshold Circuits Using 3D IC Technology

Sandeep Kumar Samal[†], Yang Li, GuoQing Chen[§], and Sung Kyu Lim[†]
[†]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
[§]Advanced Micro Devices, Beijing, China
sandeep.samal@gatech.edu, limsk@ece.gatech.edu

***Abstract*—Near-threshold computing (NTC) circuits have been shown to offer significant energy efficiency and power benefits, but with a huge performance penalty. In this paper, we demonstrate that 3D IC technology can overcome this limitation. We present a detailed case study with a 28nm commercial-grade core at 0.6V operation optimized with various 3D IC physical design methods. Our study shows that 3D IC NTC design outperforms 2D IC NTC by 29.5% in terms of performance at comparable energy. We also achieve almost 4X energy saving compared with the nominal voltage designs.**

I. Introduction

Near-threshold computing (NTC) has been researched as one of the most attractive ways to achieve significant energy savings in current VLSI systems ranging from smart low power sensors and medical devices to high performance servers. However, excessive performance degradation has prevented the use of NTC in practical applications. On the other hand, the advent of 3D IC technology has opened up a completely new design exploration space for integrated circuits.

NTC and 3D IC provide mutual benefits to bring the best out of both. While NTC designs have an order of magnitude lower power resulting in reduced thermal problems and power delivery demand, 3D ICs help in improving the performance both at the physical design and architecture levels. Architecture level synergistic benefits have been demonstrated in earlier works [1] but none of the earlier works have studied the impact of 3D IC physical design itself on full chip performance boost. In our work, we demonstrate 29.5% NTC performance improvement in 3D IC with similar energy as 2D by using various design techniques.

II. NTC Design Implementation

We use 28nm technology with multi-V$_{TH}$ library for our full RTL to GDSII block-level implementation of an OpenSPARC T2 single core design. We design and compare 2D IC at nominal (1.05V) and near-threshold (0.6V) voltages and 2-tier Through Silicon Vias (TSV) based 3D IC at 0.6V. All the designs are pushed till maximum achievable frequency of operation with no timing violation on any path.

A. Full-chip 2D/3D NTC Design Flow

We used commercial standard CAD tools with the addition of a few 3D specific in-house tools for all our designs. We used Design Compiler for netlist synthesis followed by Cadence Encounter for place and route optimization. We designed the T2 core at block level based on the top level architecture. We carried out floorplanning using simulated annealing on soft blocks with area as constraint and inter-block wirelength as cost function. The block level 2D and 3D implementations with placement and routing are shown in Figure 1. We determined timing budget of blocks in a top-down approach and then designed the individual blocks based on these timing constraints at their I/O pins.

TABLE I
Summary of the three different implementations of OpenSPARC T2 single-core. The number in brackets denote the percentage of total cell count to the nearest integer.

	Nominal 2D IC	NTC 2D IC	NTC 3D IC
Footprint (mm^2)	1.64 x 1.75	1.64 x 1.75	1.20 x 1.20
Max Frequency (MHz)	813.0	**116.3**	**150.6**
Cell Count (x1000)	365.7	366.8	386.4
Buffer Count (x1000)	53.9 (15%)	54.7 (15%)	64.5 (17%)
HVT Cells (x1000)	278.6	253.5	257.6
RVT Cells (x1000)	71.7 (20%)	103.7 (28%)	102.9 (27%)
LVT Cells (x1000)	15.4 (4%)	9.6 (3%)	25.9 (6%)
Wirelength (m)	14.8	14.6	14.7

B. 3D NTC Design with Block Folding

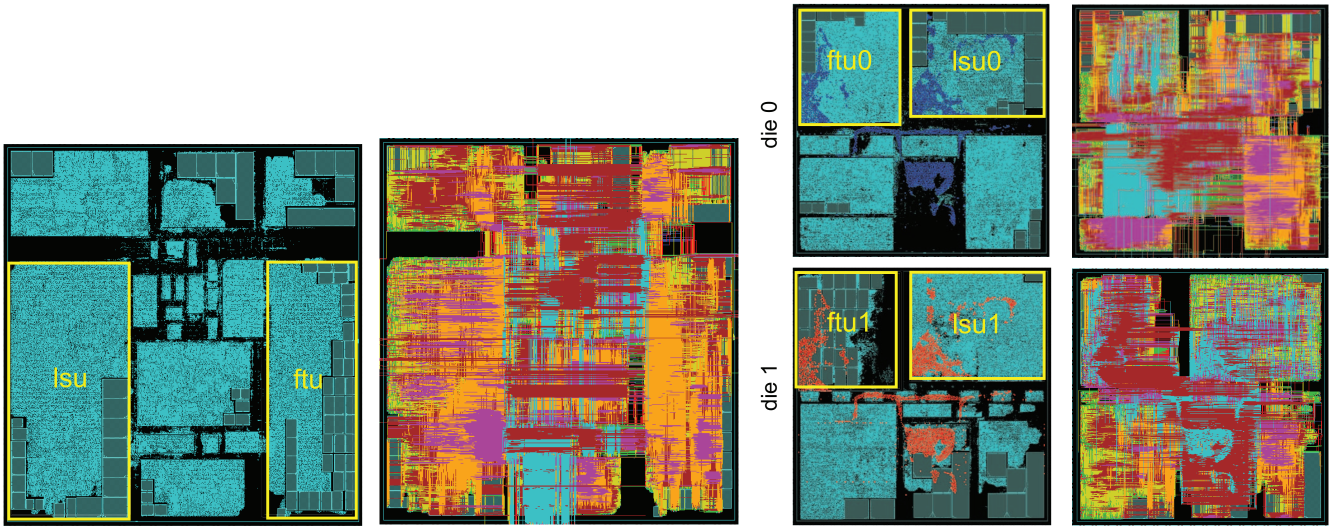
While multi-V$_{TH}$ optimization helps in improving speed in 2D OpenSPARC T2, the presence of long nets affects the overall timing and also increases power due to increased wirelength. 3D implementation facilitates shortening of nets in general. To reduce the net lengths further, we implement a two stage design folding strategy [2]. First, we select the most power hungry blocks in the design and fold them into two tiers. The folding is carried out based on the intra-block architecture such that the highly connected sub-modules remain in the same tier. These folded blocks have their own intra-block 3D TSV connections and communicate with the other blocks in the design through their block pins similar to the 2D IC implementation. TSVs are 4μm diameter with R=40mΩ and C=10fF

Based on this folded netlist of the blocks, we carry out top level partitioning and 3D floorplanning to reduce the intra-block wirelength. Using the 3D folding results and die location of the blocks, we use the netlist connectivity in each die to partition the pins of the folded blocks (lsu and ifu_ftu) into the two separate dies. This pin partitioning strategy not only ensures reduced wirelength and enhanced connectivity, but also reduces the addition of too many TSVs. All block pins are placed at the boundaries of the respective blocks. For the folded blocks, the internal TSV locations are inside the block area. The top level TSVs are only placed at inter-block whitespace. Another important design feature is the intentional use of large white space between blocks in die0 to facilitate optimized TSV insertion and ensure short connections between blocks. However, in the process of allocating white space, we maintain the overall silicon area to be the same in 2D and 3D implementations (Table I).

III. 3D NTC Performance Boost

All our designs are targeted to achieve maximum attainable frequency. We observe that nominal 2D IC reaches up to 813 MHz (1.23ns clock) while the best frequency of NTC 2D IC is 116.3 MHz (8.6ns clock). 2-tier NTC 3D IC on the other hand beats its 2D counterpart by a significant margin of 29.5% by going up to a frequency of 150.6 MHz (6.64ns clock) (Table I).

978-1-5090-0259-7/15/$31.00 ©2015 IEEE



(a) NTC (0.6V) OpenSPARC T2: 2D IC layouts

(b) NTC (0.6V) OpenSPARC T2: 3D IC layouts

Fig. 1. Near-V_{TH} (Vdd = 0.6V) OpenSPARC T2 single-core layouts. (a) 2D implementation (footprint 1.75x1.64mm), (b) 3D implementation (footprint = 1.2x1.2mm). Folded blocks (lsu and ftu) are highlighted in yellow. There are 3381 TSVs shown in blue in die0 and the corresponding landing pads are in red in die1 in the placement view. Top-level, lsu, and ifu_ftu have 1531, 1132, and 718 TSVs respectively. All layouts are shown to scale.

TABLE II
POWER-PERFORMANCE COMPARISON. NUMBERS IN BRACKETS DENOTE PERCENTAGE RELATIVE TO NOMINAL 2D.

	Nominal 2D	Near-V_{TH} 2D	Near-V_{TH} 3D
Frequency (MHz)	813.0	116.3 (14%)	150.6 (19%)
Switching Power (mW)	224.3	9.7 (4%)	9.9 (4%)
Internal Power (mW)	633.2	23.9 (4%)	31.6 (5%)
Leakage Power (mW)	16.7	1.2 (7%)	1.4 (8%)
Total Power (mW)	874.2	34.8 (4%)	42.9 (5%)
Power Delay Product (pJ)	1075.3	299.3 (28%)	284.9 (27%)

It is interesting to note that 3D IC has more cells compared to its 2D counterpart at 0.6V. This is because it is possible to insert more buffers in the 3D design to achieve faster clock periods without extra power overhead as the nets are quite short. Short nets result in shorter transition times and lower switching power per net. On the other hand, 2D design has long nets which cannot be optimized even with increased buffer count. Cadence-Encounter modifies the netlist during pre-CTS optimization based on timing and power constraints. Buffers are added, and the type and count of cells change, e.g., a multi-input AND is replaced with multiple 2-input ANDs. Timing is successfully closed for 3D IC at a faster clock compared to 2D IC and there are more such netlist changes for 3D IC. Therefore, 3D IC designs contain more cells apart from extra buffers.

Table II shows the results of post-layout power and timing analysis carried out with Synopsys PrimeTime. Primetime reports internal power which is dynamic plus short-circuit power inside standard-cells due to switching of internal-nodes only. Our 3D IC design has more cells due to tighter clock constraints, more low-Vth cells, and run 29.5% faster. Therefore, internal power is higher. However, 3D inter-cell net-switching power is still similar to 2D because of shorter nets, i.e. lower capacitive load. There are 405.6K nets in 3D and 383.6K nets in 2D design at 0.6V but the overall wirelength is almost equal which implies that the average net length is shorter in 3D IC. More LVT cells in 3D IC result in higher leakage but helps in getting the performance boost. The scaling of voltage in 2D IC domain reduces power by 25X and performance by 7X, resulting in power-delay product (PDP) savings of 3.6X. NTC 3D IC not only increases performance by 29.5% over NTC 2D IC, but also reduces PDP by another 5%.

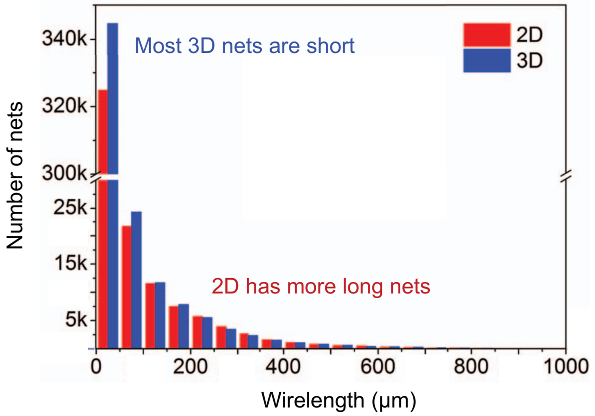


Fig. 2. Number of nets in different wirelength bins for NTC implementation with 2D and 3D.

Figure 2 shows the distribution of nets in both 2D and 3D designs at NTC based on their physical lengths. We clearly see that 3D nets are mostly shorter in length and fall in the minimum bin of distribution. Although the number of nets is higher in 3D due to more cells, their short length has lesser load capacitance and does not degrade the transition time of signal seen by the next cells in the paths.

IV. CONCLUSION

We improve NTC performance by 3D IC physical design using block folding and pin partitioning and demonstrate 29.5% faster performance than 2D IC NTC design with similar energy for OpenSPARC T2 single core processor.

REFERENCES

[1] D. Fick, *et al*, "Centip3De : A 3930 DMIPS/W Configurable Near-Threshold 3D Stacked System with 64 ARM Cortex-M3 Cores," in *ISSCC Dig. Tech. Papers*, 2012, pp. 190–191.

[2] M. Jung, *et al.*, "How to reduce power in 3D IC designs: A case study with OpenSPARC T2 core," in *CICC*, Sept 2013, pp. 1–4.