Thermal Issues in Deep Sub-Micron FDSOI Circuits

Can Baltacı, Yusuf Leblebici
Microelectronic Systems Laboratory
École Polytechnique Fédérale de Lausanne
Email: can.baltaci@epfl.ch, yusuf.leblebici@epfl.ch

***Abstract*—Self-heating effects became more prominent with the introduction of the modern devices like FDSOI and low thermal conductivity materials such as SiO_2. Consequently, the design of high speed digital circuits which are the time-critical blocks of high performance processors started to be limited mainly by thermal issues. For observing the thermal behaviour of FDSOI structure on circuit level, a 64-bit Kogge-Stone parallel prefix adder is designed and implemented in 40nm bulk CMOS technology and thermal model of the circuit is extracted and simulated according to FDSOI design parameters. The implemented adder circuit has a critical path delay of 148ps under 900 mV power supply voltage with a power consumption of 12mW. The temperature profile of the designed circuit is extracted with thermal simulations and the peak temperature locations are examined in detail. The hot spot locations and their temperature values are correlated with the power density. It is shown that self-heating of high power density devices has a significant influence on the peak temperature of a design. Finally, a simple design solution is proposed which can significantly decrease the peak temperature.**

***Keywords*—High speed digital, FDSOI, self-heating, thermal simulations, reliability**

I. Introduction

The demand for increasing the performance of high speed digital circuits brings the need for smaller and faster implementations [1]. However, as the clock frequencies increase owing to smaller technology nodes, the circuits consume power in higher densities. Consequently, the temperature levels are elevated and the thermal issues become the bottleneck of the circuits by altering the performance and decreasing the lifetime. On the performance side, the temperature induced reduced mobility decreases the device current and the maximum speed of operation [2]. Moreover, the threshold voltage decreases with the temperature and this results in higher leakage current and power consumption [3]. Higher power consumption brings higher temperature and this might result in thermal runaway where the die fails due to the uncontrolled increase in the temperature. In case thermal runaway does not happen, the chip might settle down to a higher temperature, which will degrade the performance as well as the reliability of the chip [4]. Electromigration phenomena is another reliability problem related to temperature where the metal interconnects are broken due to diffusion or flow of atoms under very high current densities at high temperatures [5]. All of the mentioned problems show that having a reliable and high performance chip is not possible without considering the thermal behaviour of the design. This brings another aspect into the design space, which is the self-heating.

Self-heating became a more critical problem especially after the introduction of the modern MOSFET device geometries like FinFET and Fully Depleted Silicon on Insulator (FDSOI) [6]. Previously, it was reported that the peak temperature of the FDSOI devices is located close to the drain end of the device [1], [7] and the peak temperature value in FDSOI FETs is found to be much higher than the one in the conventional bulk MOSFETs [8]. The higher peak temperature of the FDSOI structure is mainly due to the thermal behaviour of its constitutive materials. The thermal conductivity of the SiO_2 isolation layer is two orders of magnitude lower than the thermal conductivity of the bulk Si. Moreover, the thermal conductivity of the Si thin film, where the devices are generating heat, is one order of magnitude less than the thermal conductivity of bulk Si [1]. Additionally, the boundary between Si and SiO_2 creates a temperature jump, hence a finite interface thermal resistance, [9], which is equal to the thermal resistance of a SiO_2 layer with a thickness of 20nm [10]. Due to the mentioned facts, the dissipated power in FDSOI devices does not find a high conductance diffusion path. Consequently, the generated heat turns into temperature in nanometer scale local spots which are comparable to the dimensions of transistors in FDSOI. However, not all the devices settle down to very high temperature values in an implementation. The devices which consume the highest amount of power per unit area are the hottest ones especially in FDSOI. As a result of this, a design which contains devices with large differences in their power densities create very high temperature hot spots and large temperature gradients. By performing a detailed power density analysis, the critical ones can be eliminated from the others; and by performing some modifications on their design, the peak temperature and the high temperature gradients can be reduced. Recently, during the implementation of a 5 GHz processor, high switching factor nets were identified during functional simulation to avoid micro hot-spots at the individual gate level, caused by device self-heating [11]. As a solution, the maximum output load capacitances of the gates driving these nets are reduced and these gates are placed away from the other gates driving such nets in order to avoid excessive heating and have uniform temperature distribution overall the circuit.

In this work, it is intended to emphasize the correlation of the nanometer scale hot spots and the power density of individual devices by observing the temperature profile of high performance circuits. For that, a 64-bit parallel prefix adder is designed and implemented in a commercially available 40

978-1-5090-0493-5/16/$31.00

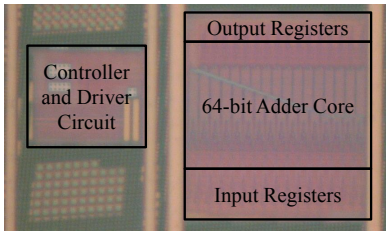
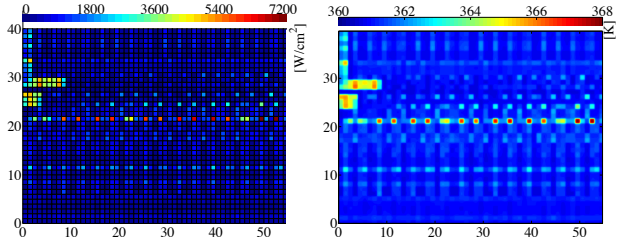
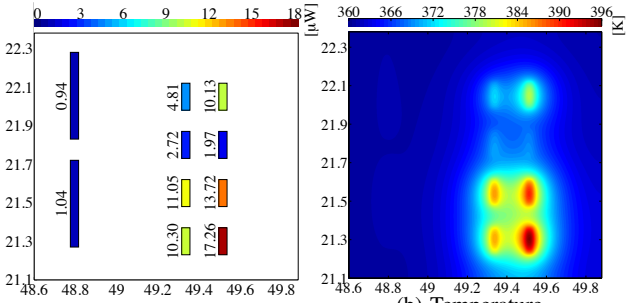


Fig. 1: Die micrograph of the implemented 64-bit adder. The thermal simulations of Fig. 2 are performed on the 64-bit Adder Core part.



(a) Heat generation (b) Temperature

Fig. 2: Heat generation and the resulting temperature map of the 64-bit adder (x and y dimensions are in μm).



(a) Heat generation (b) Temperature

Fig. 3: Heat generation input of a high resolution thermal simulation and the resulting temperature map focused on a small portion of the 64-bit adder where the heat generation is very high (x and y dimensions are in μm).

nm CMOS bulk technology. The power dissipation of each device in the circuit is observed under randomly applied input vectors. The resulting power dissipation output is provided as an input to thermal simulations for observing the temperature profile of the overall block. HotSpot tool [12] is used for modelling the thermal behaviour of the FDSOI geometry. For that, the technology parameters of a commercially available 28nm FDSOI design kit are used by scaling them for 40nm. The devices situated on the highest temperature locations are found and examined. It is observed that the local hot spots have sizes comparable to the size of the devices and the generated heat is directly converted to temperature in FDSOI. Consequently, the highest temperature values occur on the devices which has the highest power density. Finally, a solution for decreasing the temperature of the hotspots is proposed. It is shown that the peak temperature of the deign in FDSOI can be decreased significantly with a cost of an insignificant increase in the area and parasitic capacitances.

In Section II, the performance parameters of the implemented 64-bit parallel prefix adder are given. In Section III, the thermal simulation results and the temperature profile of the designed 64-bit adder is provided. Finally, in Section IV, the summary of the work and the conclusions are provided.

II. Implemented Block

The parallel prefix adder is implemented with Kogge-Stone technique [13] where radix-4 and sparseness-4 options are used [14]. The entire 64-bit Kogge-Stone adder block is designed with full custom design approach (Fig. 1). The block is primarily optimized to obtain the lowest possible critical path delay while having the lowest possible power consumption and area. Finally, a delay (clock to Sum) of 148ps is obtained under 900 mV power supply voltage. The block contains 10922 nMOS and pMOS devices and the resulting area of the block is around $2200\mu\mathrm{m}^2$. The average power dissipation of the block is 12mW and the average power density is $548\mathrm{W}/\mathrm{cm}^2$.

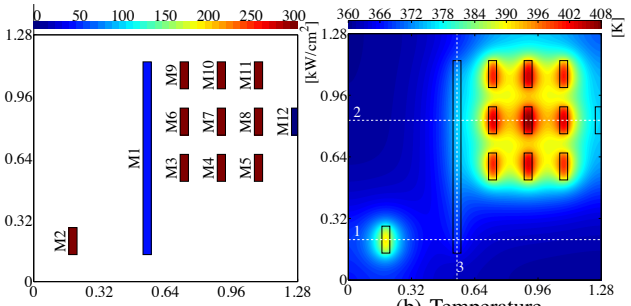
III. Self-Heating Analysis

For extracting the power dissipation waveforms of each device in the block, series of randomly generated input vectors are applied to the inputs. The instantaneous power dissipation waveforms of each device is extracted by (1) and the average power dissipation values are calculated by (2) where P_n is the instantaneous power dissipation (in Watts), $I_{D,n}$ is the drain current (in Amperes), $V_{DS,n}$ is the voltage drop between the drain and the source terminals (in Volts), $\overline{P_n}$ and $\overline{Q_n}$ are the average power dissipation and heat generation (in Watts) respectively of device n throughout the simulation time (t_s, in seconds).

$$P_n(t) = I_{D,n}(t) \times V_{DS,n}(t) \tag{1}$$

$$\overline{P_n} = \frac{1}{t_s}\int_0^{t_s} I_{D,n}(t) \times V_{DS,n}(t)\mathrm{d}t = \overline{Q_n} \tag{2}$$

The locations of each device are extracted from layout for performing thermal simulations. During the thermal simulations of the entire adder circuit, the block is partitioned into squares (pixels) of $1\mu\mathrm{m}^2$ and the total heat generation of each square is calculated by summing the power dissipation of each device which falls in that square (Fig. 2a). The resulting temperature map (Fig. 2b) shows that the generated heat is locally translated into temperature since the thermal conductance of the FDSOI structure is quite low at the heat generation spots. This is mainly due to the SiO_2 isolation layer, Si-SiO_2 thermal boundary resistance [9] and the reduced thermal conductance of silicon and SiO_2 thin films [15]. For observing the temperature profile of individual devices and the hottest spot in the entire block, it is needed to further increase the resolution of the simulation. For that, the location of the pixel with the highest temperature is determined on Fig. 2a and the thermal simulations are repeated with higher



(a) Heat generation (b) Temperature

Fig. 4: Heat generation of the test case and the resulting temperature profile (x and y dimensions are in μm).

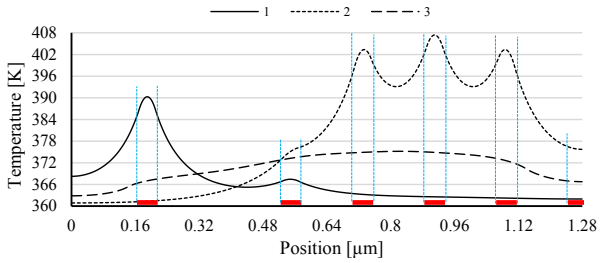


Fig. 5: Temperature profiles along the cut-lines (Fig. 4b) taken at the center of the devices M2 ($y = 1.07\mu$m), M3 ($y = 0.64\mu$m) and M4 ($y = 0.21\mu$m). The red thick lines show the location of the devices on Fig. 4.

accuracy. Fig. 3a shows the devices which are located on the spot with the highest temperature and their power dissipation values. The resulting temperature profile (Fig. 3b) indicates that the generated heat does not diffuse easily and each device heats up itself significantly. The temperature difference inside this small window ($1.6\mu m^2$) is larger than 35K and a temperature gradient of more than 20K/μm can be observed. Moreover, without being too obvious, the same picture implies that a device with very high power dissipation might have a considerable impact on the temperature of its neighbouring device (observable from the four high power devices located on the lower right part of Fig. 3b).

To understand the effect of neighbour devices on each other, a heat generation scenario is created. Fig. 4a shows the locations of the devices with their names indicated on the left. The colours of the devices are determined according to the colour-bar on the top, which represents the average power density. The instantaneous and average power density of each device is defined by (3) and (4) respectively, where W_n and L_n are the width and length of the corresponding device:

$$PD_n(t) = \frac{P_n(t)}{W_n \cdot L_n} \tag{3}$$

$$\overline{PD_n} = \frac{\overline{P_n(t)}}{W_n \cdot L_n} \tag{4}$$

For the test case of Fig. 4a, the sizes of the devices are selected according to the design rules of the utilized technology and the power density values are chosen according

TABLE I: Parameters of the devices on Fig. 4

Device	M1	M2 - M11	M12
Power Density [kW/cm^2]	42	300	0
Power Dissipation [μW]	16.8	16.8	0
Width [μm]	1.00	0.14	0.14
Length [μm]	0.04	0.04	0.04

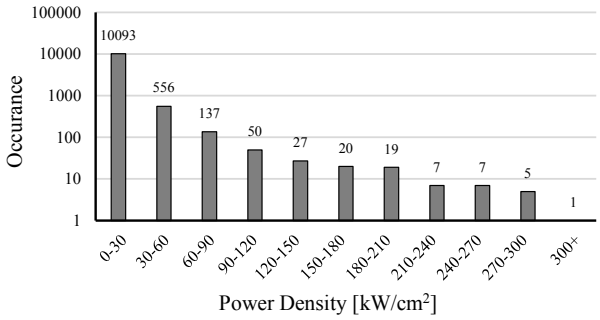


Fig. 6: Power density distribution of the transistors which belong to the 64-bit adder circuit.

to the power density distribution of the adder circuit (Fig. 6). The power density of devices M2-M11 is equal to the power density of the device which has the highest power density in the entire adder circuit according to the power dissipation analysis. Naturally, one of these devices should be responsible of the peak temperature of the overall implementation. On the other hand, M12 has the same size as M2-M11 with no power dissipation. Finally, the power dissipation of M1 is equal to the power dissipation of M2-M11 while its power density is much lower than the others due to its larger size. The horizontal distance between the devices is set to the minimum gate-to-gate distance permitted by the used design kit. All of the other parameters of each device are summarized on Table I.

The resulting temperature profile of the listed heat generators is shown on Fig. 4b. The white dashed lines on Fig. 4b are the cut-lines which pass on the center of the devices. The temperature waveforms occurring on these cut-lines are shown on Fig. 5. The peak temperature on cut-line 1 shows that the device with the highest power density (M2) can increase the temperature itself alone by 30K. The peak temperature on cut-line 2, which is the peak temperature of the entire area, is observed on M7 with a temperature increase of more than 48K when compared to the coolest point. Additionally, the temperature of M7 is 18K higher than the temperature of M2 although they have the same heat density. This proves that the influence of the neighbouring devices (M3-M6 and M6-M11 in the example) can increase the peak temperature of the overall circuit significantly. Another observation which can be made on cut-line 2 is although M12 does not dissipate any power, its temperature is 15K higher than the lowest temperature in the area. This is primarily due to the self-heating of M3-M11 which are located very close to M12. The temperature of M1 is 15K lower than the temperature of M2 although their power

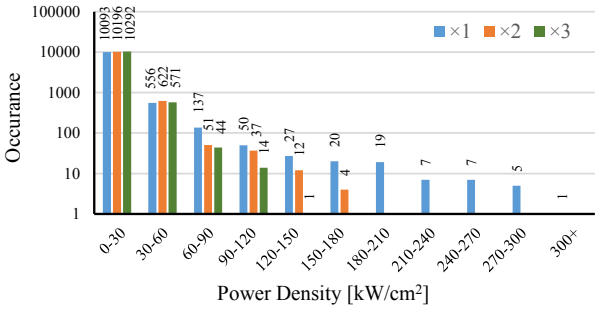


Fig. 7: The new power density distribution histograms after increasing the width of the high power density devices by a factor of 2 and 3.

dissipation values are equal. This result proves that is the most substantial factor in terms of creating a nanometer scale hot spot in FDSOI is the power density rather than the power dissipation alone. Finally, the temperature of M1 shows more than 6.8K variation along its width (cut-line 3 on Fig. 5). The gradient is due to the neighbouring devices (M3, M6, M9) which are very strong heaters. As a result of this variation, the threshold voltage and mobility of M1 would show different behaviours along its width. While this effects only the speed in digital circuits, it might create important mismatch problems in analog circuits.

The analysis shows that the hottest spot in FDSOI is most probably created by the self-heating of individual devices with the highest heat density. Moreover, their cumulative effect can increase the peak temperature significantly. Returning back to the 64-bit adder circuit, it can be seen that the power density range is limited by $[0\mathrm{kW/cm^2}, 310\mathrm{kW/cm^2}]$ window and 99% of all the devices has a power density value less than $100\mathrm{kW/cm^2}$ (Fig. 6). This means that only 1% of the devices are generating more than 33% of the maximum power density observed in the entire circuit. Therefore, if the devices in this 1% portion are detected and modified in a way, the peak temperature of the block can be significantly decreased. The modification of the most critical devices with high power density can be performed by increasing their width with a trade-off of a slightly increased area and parasitic capacitances. Fig. 7 shows that the power density distribution can be squeezed into a smaller window only by increasing the width of the devices with power density of more than $100\mathrm{kW/cm^2}$. The new maximum power density values are less than $170\mathrm{kW/cm^2}$ and $125\mathrm{kW/cm^2}$ respectively for the $\times 2$ and $\times 3$ cases (Fig. 7). Another improvement can be provided by separating the high power density devices from each other to relax the heat flow [11]. Finally, thermal vias can be added to the drain ends of these devices similar to what is shown in [16] and [17] to provide better heat diffusion paths, which will in turn increase the delay time.

IV. Conclusion

In this paper, the thermal problems of very high speed digital blocks in FDSOI are studied by experimenting a 64-bit parallel prefix adder operating under 900 mV power supply voltage. Thermal simulations on FDSOI show that the devices with high power density values are the main factors of the hot spots in a circuit due to the poor thermal conductance of FDSOI geometry. A single device with practical dimensions and power dissipation values can be more than 30K hotter compared to its surroundings only due to its self-heating. Moreover, closely placed devices can increase the temperature of their neighbours by 20K. Finally, it is demonstrated that sizing the critical devices slightly bigger would decrease the maximum power density by 60% with a trade-off of a slight decrease in the speed and an increase in the area.

References

[1] K. R. Dragica Vasileska and S. M. Goodnick, "Heating effects in nanoscale devices," in *Cutting Edge Nanotechnology*, D. Vasileska, Ed. InTech, 2010, ch. 3, pp. 33–60.

[2] Y. Tsividis, *Operation and Modeling of the MOS Transistor*. New York, NY, USA: McGraw-Hill, Inc., 1987.

[3] S. S. Bhattacharya, S. K. Banerjee, B. Y. Nguyen, and P. J. Tobin, "Temperature dependence of the anomalous leakage current in polysilicon-on-insulator mosfet's," *IEEE Transactions on Electron Devices*, vol. 41, no. 2, pp. 221–227, Feb 1994.

[4] M. L. Mui, K. Banerjee, and A. Mehrotra, "Power supply optimization in sub-130 nm leakage dominant technologies," in *Quality Electronic Design, 2004. Proceedings. 5th International Symposium on*, 2004, pp. 409–414.

[5] J. R. Black, "Electromigration-a brief survey and some recent results," *IEEE Transactions on Electron Devices*, vol. 16, no. 4, pp. 338–347, Apr 1969.

[6] E. Pop and K. E. Goodson, "Thermal phenomena in nanoscale transistors," in *Thermal and Thermomechanical Phenomena in Electronic Systems, 2004. ITHERM '04. The Ninth Intersociety Conference on*, June 2004, pp. 1–7 Vol.1.

[7] E. Pop, S. Sinha, and K. E. Goodson, "Heat generation and transport in nanometer-scale transistors," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1587–1601, Aug 2006.

[8] C. Ni, Z. Aksamija, J. Y. Murthy, and U. Ravaioli, "Coupled electro-thermal simulation of mosfets," *Journal of Computational Electronics*, vol. 11, no. 1, pp. 93–105, 2012.

[9] E. Lampin, Q.-H. Nguyen, P. A. Francioso, and F. Cleri, "Thermal boundary resistance at silicon-silica interfaces by molecular dynamics simulations," *Applied Physics Letters*, vol. 100, no. 13, 2012.

[10] E. Pop, "Self heating and scaling of thin body transistors," Ph.D. dissertation, -Stanford University, 2005.

[11] J. Warnock et al, "4.1 22nm next-generation ibm system z microprocessor," in *Solid- State Circuits Conference - (ISSCC), 2015 IEEE International*, Feb 2015, pp. 1–3.

[12] W. Huang, K. Skadron, S. Gurumurthi, R. J. Ribando, and M. R. Stan, "Differentiating the roles of ir measurement and simulation for power and temperature-aware design," in *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on*, April 2009, pp. 1–10.

[13] P. M. Kogge and H. S. Stone, "A parallel algorithm for the efficient solution of a general class of recurrence equations," *IEEE Transactions on Computers*, vol. C-22, no. 8, pp. 786–793, Aug 1973.

[14] Y. Shimazaki, R. Zlatanovici, and B. Nikolic, "A shared-well dual-supply-voltage 64-bit alu," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 3, pp. 494–500, March 2004.

[15] T. Yamane, N. Nagai, S.-i. Katayama, and M. Todoki, "Measurement of thermal conductivity of silicon dioxide thin films using a 3 method," *Journal of Applied Physics*, vol. 91, no. 12, 2002.

[16] T.-Y. Chiang, K. Banerjee, and K. C. Saraswat, "Effect of via separation and low-k dielectric materials on the thermal characteristics of cu interconnects," in *Electron Devices Meeting, 2000. IEDM '00. Technical Digest. International*, Dec 2000, pp. 261–264.

[17] Z. Wang, G. Dong, Y. Yang, and J. Li, "Effect of dummy vias on interconnect temperature variation," *Chinese Science Bulletin*, vol. 56, no. 21, pp. 2286–2290, 2011.