A Thermally-Aware Energy Minimization Methodology for Global Interconnects

Soheil Nazar Shahsavani, Alireza Shafaei, Shahin Nazarian, and Massoud Pedram
Department of Electrical Engineering, University of Southern California
Los Angeles, CA 90089 USA
nazarsha@usc.edu, shafaeib@usc.edu, shahin@usc.edu, pedram@usc.edu

***Abstract*—As a result of the Temperature Effect Inversion (TEI) in FinFET-based designs, gate delays decrease with the increase of temperature. In contrast, the resistive characteristic and hence delay of global interconnects increase with the temperature. However, as shown in this paper, if buffers are judiciously inserted in global interconnects, the buffer delay decrease is more pronounced than the interconnect delay increase, resulting in an overall performance improvement at higher temperatures. More specifically, this work models the delay of buffer-inserted global interconnects vs. temperature in order to derive the optimal number and size of buffers for a given interconnect length and temperature. Furthermore, the paper addresses the problem of minimizing the buffered interconnect energy consumption by changing the supply voltage level or FinFET threshold voltage, and also presents a temperature-aware optimization policy for solving this problem. Simulation results show average interconnect energy savings of 16% with no performance penalty for five different benchmarks implemented on a 14nm FinFET technology.**

I. Introduction

The demand for higher computational speed as well as lower power consumption has led to a rapid downscaling of technology. However, due to serious drawbacks of bulk CMOS technology especially in sub-20nm nodes (e.g., poor control over the channel, high sub-threshold leakage which leads to higher power consumption, and device-to-device variation), *double-gate field-effect transistors (DG-FETs)* have been recognized as a promising replacement for MOSFET. This is because of the superior control of *short channel effects* (SCEs), lower power consumption, and better voltage scalability of DG-FET devices. Among different implementations of DG-FETs, FinFETs are one of the most compelling options due to higher immunity to random dopant variations and soft errors as a result of the undoped channel of FinFET devices [1][2].

Accordingly, industry has begun replacing planar CMOS transistors with quasi-planar FinFET devices mainly because of the improved (three-dimensional) gate control over the channel, which in turn diminishes source and drain controls, thereby reducing SCE [3]. Additionally, the minimum energy point and the minimum energy-delay point of FinFET circuits occur at supply voltage, V_{DD}, levels lower than that of planar CMOS counterparts [4], enabling more aggressive voltage scalability in FinFET-based circuit designs. Thanks to higher ON/OFF current ratio and better controllability over the threshold voltage, FinFET devices are significantly more power efficient than the equivalent planar CMOS process. Hence, FinFETs are widely recognized as the technology-of-choice beyond the 20nm regime [5].

Technology shrinking has also significantly increased the power density, which in turn leads to a higher rate of heat generation per chip area and therefore more thermal hotspots. These thermal hot spots along with high temperature gradients, lead to reliability issues and performance degradation [6]. Failure mechanisms including electromigration, self-heating, and time-dependent-dielectric-breakdown, accompanied by super exponential increase in leakage power due to a positive feedback between temperature and leakage current, enforces exploitation of both power and thermal management techniques [7].

Interconnect delay has become a major concern for the overall chip performance, as the chip performance is being aggressively dominated by the delay of global and semi-global interconnects [8]. The resistive characteristic of metal interconnects exacerbates this situation, since the electrical resistance increases linearly with the temperature, which in addition to quadratic dependency of delay on the wire length, downgrades the overall performance significantly. Furthermore, it has been estimated that more than 50% of the total power consumption of multiprocessors is dissipated by repeaters in the interconnect network, dominating the power consumption of logic gate [9].

In addition to previously mentioned attributes of FinFET devices, one of the most interesting characteristics of such devices is that the gate delay decreases with the increase of temperature in all the operation voltage regimes, from super-threshold to sub-threshold [7]. This phenomenon, called *temperature effect inversion* (TEI), could lead to considerable performance improvement and power reduction in digital designs. This is especially useful for high performance applications where one could take advantage of intrinsically high die temperature of state-of-the-art integrated circuits to greatly enhance the performance or lower the power consumption.

To the best of our knowledge, no previous work has been dedicated to energy optimization of global interconnects in FinFET devices considering the TEI phenomenon. The key contributions of this paper could be summarized as follows:

- We analytically derive the delay-optimal buffer size and wire segment length of buffer-inserted global interconnects as a function of the temperature.
- We present a temperature-aware energy minimization policy with no performance overhead for interconnects. Our results show on average 16% interconnect energy saving with no performance penalty for five randomly chosen benchmarks from SPLASH2 benchmark suit [10], under a 14nm FinFET technology.

The rest of the paper is organized as follows. Section II overviews the related work. In Section III, basic concepts including the TEI phenomenon are expressed. The buffer insertion and energy minimization techniques are presented in Section IV. Section V provides simulation results. Finally, the paper is concluded in Section VI.

II. Related Work

There has been a wide range of modeling and optimization techniques for interconnect delay and power consumption. A power-optimal repeater insertion technique for global interconnects is proposed in [11], which reduces the power dissipation

with a limited performance penalty by changing the optimal buffer size. A wire and buffer sizing algorithm based on Lagrangian relaxation is introduced in [12] to simultaneously optimize delay, power, skew, and area in clock tree designs. Ref. [13] presents optimal algorithms for timing optimization and minimization of dynamic power dissipation and area by discrete wire sizing. In [14], a timing model based on the short-channel α-power law model is presented in order to determine the optimum number of uniformly sized repeaters along a resistive interconnect line to reduce the delay. This method also compares the short-circuit power and dynamic power dissipation in repeater chains and discusses the related power/delay trade-offs. Authors in [15] introduce a methodology for minimizing power and area overheads of repeaters while meeting the target performance goals by integrating area and power overhead constraints along with delay into the repeater design methodology. Also, a method for threshold voltage control using multiple supply voltages for power-efficient FinFET interconnects has been proposed in [16].

III. Preliminaries

A. Temperature Effect Inversion Phenomenon

Propagation delay of a logic gate is inversely proportional to the current of the driving transistor I_{on}. In the super-threshold regime, I_{on} can be modeled as follows:

$$I_{on} = \mu(T) \cdot C_{ox} \frac{W}{L} \cdot (V_{GS} - V_{th}(T))^{\alpha}, \tag{1}$$

where α is empirically determined for different technologies [17].

In conventional long-channel CMOS transistors, although both V_{th} and μ decrease as temperature increases, the mobility degradation effect on current dominates and results in lower I_{on} and higher propagation delay [7]. However, in FinFET devices in sub-30nm technologies, the threshold voltage effect on current dominates and leads to higher I_{on} and lower propagation delay. This effect mainly happens due to the bandgap narrowing induced by tensile stress effect of the insulator [18]. As technology scales down, the effect of the insulator induced tensile stress from the layer to the fin body changes the device characteristics more significantly due to the fact that the thinner fin body has larger stress, thus stress induced bandgap narrowing results in a more aggressive degradation of the threshold voltage in sub-30nm technologies. As the temperature increases, the tensile stress becomes larger, which decreases V_{th} more aggressively than the carrier mobility in FinFET technology. Note that, the effect of temperature on drain current of MOSFETs in sub/near-threshold regimes shows similar characteristics to that of FinFET devices operating in the super-threshold regime and below, that is, the current increases as the temperature goes up. I_{on} in sub/near threshold regimes can be approximated as [19]:

$$I_{on,subth} = I_0 \cdot e^{\frac{(V_{GS} - V_{th}(T))}{S(T)}} \cdot (1 - e^{-\frac{V_{DS}}{\nu_T}}) \tag{2}$$

where I_0 denotes the current at $V_{GS} = V_{th}$, and is dependent on process, device geometry, temperature and mobility. $\nu_T = \frac{KT}{q}$ is the thermal voltage (K is the Boltzman constant and q denotes the electron charge). $S(T) = n\nu_T$ is the subthreshold slope, defined as the change in V_{GS} to alter the drain current by an order of magnitude [20] and n is a numerically determined parameter.

In the sub/near-threshold regime, V_{th} reduction dominates the mobility degradation, and hence, due to the exponential dependency of drain current on the threshold voltage, a higher I_{on} value is achieved in higher temperatures [21][22]. Threshold voltage degradation in conjunction with the tensile stress effect, cause a significant delay reduction in FinFETs operating in the sub/near-threshold regimes as the temperature goes up. Consequently, as a result of increasing the temperature, the drain current of FinFETs increases in all supply voltage regimes, which results in performance enhancement.

The temperature dependency of threshold voltage of FinFET devices may be modeled as follows [23]:

$$V_{th} = \Delta\Phi_i - \nu_T \cdot ln(\frac{q \cdot n_i \cdot t_{Si}}{4 \cdot C_{ox} \cdot \nu_T}) \tag{3}$$

where $\Delta\Phi_i$ denotes the work function difference between the gate electrode and intrinsic silicon, n_i is the intrinsic carrier density, t_{Si} denotes the silicon film thickness, and C_{ox} is the oxide capacitance. This temperature dependency, derived by HSPICE simulations, and using the BSIM-CMG [24] model, may be formulated as:

$$V_{th}(T) = -\eta T + V_{th_0} \tag{4}$$

where V_{th_0} represents the threshold voltage at ($T = 0^0C$), and η is a technology dependent positive coefficient which has been calculated for BSIM-CMG model for different technology nodes. It is observed that as the channel length decreases, sensitivity of threshold voltage to temperature is increased due to higher tensile stress on the channel. As a consequence of the TEI phenomenon, the effective resistance of a single transistor decreases as temperature rises. Lower channel resistance in turn leads to lower gate delay and higher performance. Simulation results for different V_{DD} values and technology nodes for a FinFET-based 17-stage inverter chain, shown in Fig. 1, validates the theory.

B. Temperature-dependent Interconnect Delay Model

Both resistance and capacitance of the VLSI interconnect increase with wire length L. Therefore, the RC delay of a wire increases with L^2. The delay may be reduced significantly by splitting the line into segments, and inserting buffers between these segments. Buffer insertion technique alleviates the quadratic increase in propagation delay, while lowering power dissipation by decreasing the short-circuit current [14]. Consider a uniform interconnect with resistance and capacitance per unit length of r_{wire} and c_{wire}, respectively, buffered by identical repeaters of size s as shown in Fig. 2. It should be noted that the effect of line inductance on the delay of the interconnect segment has been neglected in this analysis, due to the fact that line inductance reduces with technology scaling and line widths should be increased by a large factor (16×) before inductive effects become important [11].

Assuming equal rise and fall delays for repeaters, we suppose a unit-size inverter is modeled by a resistance r_{inv}, along with gate and diffusion capacitances c_g and c_p, respectively.

One could then obtain the segment delay, defined as the time difference between the input and output voltages crossing 50% of their full swing value given by $\tau ln(2)$, where the time constant τ is defined as:

$$\begin{aligned}\tau =& \frac{r_{inv}}{s} \cdot (sc_p + \frac{1}{2} c_{wire} l_{wire}) + \\ & (\frac{r_{inv}}{s} + r_{wire} l_{wire}) \cdot (\frac{1}{2} c_{wire} l_{wire} + sc_g)\end{aligned} \tag{5}$$

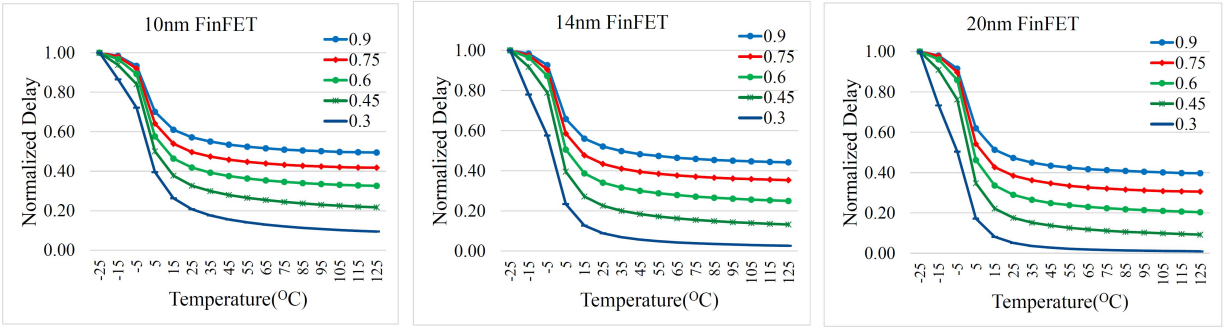

Fig. 1. Normalized gate delay at different temperatures and supply voltage levels for a FinFET-based 17-stage inverter chain.

in which $l_{wire} = \frac{L}{N}$ is the length of each wire segment between two repeaters, L denotes the overall wire length, and N is the total number of repeaters. s denotes the relative size (driver strength) of the inserted inverters compared to that of unit-size inverter. Differentiating (5) with respect to s and N yields the optimum length of wire between repeaters, l_{opt}, and the optimum size of each repeater, s_{opt}, as follows:

$$l_{opt} = \sqrt{\frac{2r_{inv}(c_g + c_p)}{r_{wire}c_{wire}}},\ s_{opt} = \sqrt{\frac{r_{inv}c_{wire}}{r_{wire}c_g}} \tag{6}$$

Substituting above values in (5), we can derive the minimum delay per unit length of an optimum length and optimally buffered (OLOB) segment as:

$$(\frac{\tau}{l})_{opt} = 2(\sqrt{r_{wire}c_{wire}r_{inv}c_{inv}}) \cdot (1 + \sqrt{\frac{1}{2}(1 + \frac{c_p}{c_g})}) \tag{7}$$

Furthermore, resistance of a wire r_{wire} is linearly proportional to temperature as follows [25]:

$$r_{wire}(T) = r_{wire}(T_0) \cdot (1 + \delta \cdot (T - T_0)) \tag{8}$$

where δ is the temperature coefficient of resistance for the conductor material, which is $0.0039\frac{1}{K}$ for copper. Based on the fact that equivalent resistance of a transistor is inversely proportional to its drain current [20], the temperature dependency of the equivalent resistance of an inverter in all regimes, r_{inv}, could be written as:

$$r_{inv}(T) \propto \begin{cases} \dfrac{1}{\mu(T)(V_{GS} - V_{th}(T))^{\alpha}} & V_{GS} > V_{th} \\ \dfrac{1}{e^{\frac{(V_{GS} - V_{th}(T))}{n\nu_T}} \cdot (1 - e^{-\frac{V_{DS}}{\nu_T}})} & V_{GS} \le V_{th} \end{cases} \tag{9}$$

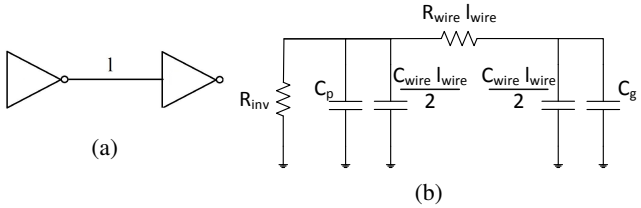

Fig. 2. An interconnect segment of length l between two identical unit-size buffers: (a) schematic representation, and (b) its equivalent RC circuit.

Finally, $r_{wire}(T_0)$ and c_{wire} are obtained using the ITRS projections of interconnects as follows [9]:

$$r_{wire}(T_0) = \frac{\alpha_{scatter} \cdot \rho_{wire}(T_0)}{(m_{thick} - t_{dish} - t_{barrier})(m_{width} - 2t_{dish})}$$
$$c_{wire} = 2\epsilon_0(\epsilon_{horiz} \cdot \frac{m_{thick}}{m_{spacing}} + \epsilon_{vert} \cdot \frac{m_{width}}{ILD_{thick}}) + c_{fringe} \tag{10}$$

where m_{thick}, $m_{spacing}$, and m_{width} denote the interconnect line thickness, spacing, and width, respectively; t_{dish} and $t_{barrier}$ are the thickness of dishing and barrier; ILD_{thick} is the interlayer dielectric thickness; ϵ_{horiz} and ϵ_{vert} represent the horizontal and vertical dielectric constant; and $\alpha_{scatter}$, ρ_{wire}, and c_{fringe} denote the scattering factor, interconnect resistivity, and fringing capacitor, respectively. In (10), ρ_{wire} is the only temperature-dependent variable.

Based on (7), as the temperature increases, r_{wire} increases and r_{inv} decreases, so $(\tau/l)_{opt}$ may increase or decrease depending on the relative degrees of change in the r_{wire} and r_{inv}. Using equations (4)(8)(9) for the threshold voltage, wire resistance and equivalent inverter resistance in the superthreshold voltage regime (neglecting the temperature dependency of mobility), and taking the derivative of (7) with respect to temperature, results in:

$$\frac{\partial(\tau/l)_{opt}}{\partial T} = c_1 \cdot \frac{r_{inv} \cdot \frac{\partial r_{wire}}{\partial T} + r_{wire} \cdot \frac{\partial r_{inv}}{\partial T}}{\sqrt{r_{wire}r_{inv}}} = \frac{c_2}{\sqrt{r_{wire}r_{inv}}} \frac{\delta(V_{DD} - V_{th_0}) + \delta\eta(1-\alpha)T - \eta\alpha}{(V_{DD} - V_{th})^{2\alpha}} \tag{11}$$

As it can be observed, based on the technology dependent parameters, α, δ, and η, and difference of supply voltage and threshold voltage at $(T = 0^0C)$, derivation could be positive, negative, or even zero, leading to a minimum point delay across different temperatures. In particular, in our adapted finFET technology, (11) is a negative number, meaning that as the temperature increases, the delay of the OLOB interconnect segment decreases. Simulation results also validate our observation.

IV. Proposed Method

A. Optimal Buffer Insertion Technique

According to (8), r_{wire} increases as temperature goes up. For instance, delay of a copper interconnect line without buffers (repeaters) increases by as much as 40% when the temperature increases from 25^oC to 125^oC. In conventional

long-channel CMOS circuits, due to mobility degradation at higher temperatures, r_{inv} also increases, and thus, the propagation delay of the interconnect increases as the temperature goes up. As a result of temperature rise, performance degradation is more rapid in a CMOS buffer-inserted line compared to a line with no buffers. However, in FinFET-based circuits, due to the TEI phenomenon, r_{inv} decreases as temperature increases, whereas r_{wire} is increased. Consequently, by selecting the proper values of s and l, the propagation delay of an OLOB segment decreases as temperature increases, and hence, interconnects speed up, reversing the general trend in interconnects.

As a result of variations of r_{inv} and r_{wire} with temperature, s_{opt} and l_{opt} decrease as temperature rises. In other words, there is a specific set of s_{opt} and l_{opt} values for each temperature. As a consequence, as temperature increases, both s_{opt} and l_{opt} values decrease, which means lower size buffers and smaller size line segments (which is equivalent to having higher number of buffers N, placed after shorter line segments in a fixed size interconnect) are needed.

Now, assume that at temperature T_1, size of the buffers s_{opt} and their spacing l_{opt} values are set to be s_1 and l_1 respectively. If the temperature is increased to T_2, the optimal delay is achieved by values s_{opt} and l_{opt} set to $(s_2 < s_1)$ and $(l_2 < l_1)$, however the size and spacing of the buffers are s_1 and l_1, so the delay would be different than the case with buffers designed for temperature T_2. Given an interconnect line at temperature (T_2) with total length of L (the least common multiple of l_1 and l_2), the delay difference in the case that buffers are optimized for T_2 compared with the case that they are optimized for T_1, may be calculated as follows:

$$\tau(s_1, N_1) - \tau(s_2, N_2) = (s_1 - s_2)(c_g r_{wire} L - \frac{c_{wire} r_{inv} L}{s_1 s_2}) + (N_2 - N_1)[\frac{c_{wire} r_{wire} L^2}{2 N_1 N_2} - r_{inv}(c_p + c_g)] \quad (12)$$

Based on values of technology dependent parameters, and temperature values T_1 and T_2, the delay difference can be higher or lower than zero. Substituting the parameters of our finFET technology model in (12), assuming $T_1 = 25^o$C and $T_2 = 65^o$C, it is inferred that the delay of the OLOB designed for temperature $T_1(s_1, l_1)$ and operating at T_2 is higher than the delay of the OLOB optimized for higher temperature of T_2 (s_2, l_2). The delay per unit length vs. temperature characteristics of an OLOB segment for different values of s (normalized by the s_{opt} value for 25^oC), as shown in Fig. 3, validates the proposed performance improvement by designing for higher temperatures. It should be mentioned that delay values have been normalized by the lowest delay at each temperature. Consequently, in order to improve the overall performance of the global interconnects, buffer insertion should be performed based on the average temperature, rather than the worst case temperature. It should be noted that unlike conventional long-channel CMOS-based designs, due to the TEI phenomenon in FinFET-based designs, the worst case delay occurs in the lowest temperature rather than highest temperature.

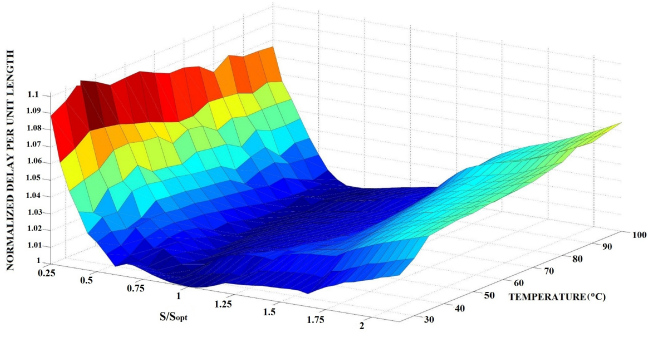

Fig. 3. Normalized delay per unit length of an OLOB segment for different buffer sizes and temperatures.

In order to obtain the temperature profile across different areas of the chip under different workloads, simulations on five benchmarks, selected from SPLASH2 benchmark suite [10] on an 8-core MPSoC using SNIPER [26] multicore simulator is performed, and temperature is calculated using Hotspot [27] tool. The ArchFP [28] was used to extract the floorplan of the MPSoC and each core is composed of one L1 private cache, one L2 private cache, one shared L3 cache as well as Instruction Fetch Unit (IFU), Execution Unit (EXU), Memory Management Unit (MMU), Renaming Unit (REU) and Load/Store Unit (LSU). Resulting temperature profiles show that different areas of the chip have different average temperatures ranging from 25^oC to 80^oC. Therefore, to capture the spatial variance in the temperature and maximize the performance improvement, different numbers of buffers of different sizes in different areas of the chip should be used.

To maximize the performance, based on the temperature profile obtained from different benchmarks, we divide the chip into the following two temperature regions: (i) a cold region dominated by cache memories where interconnects are optimized for 25^oC, and (ii) a hot region composed of datapath and register files, optimized for 65^oC. For each temperature region, optimal values of s and l are derived as discussed in the previous section.

Results of the average performance improvement of an OLOB interconnect of length 1mm optimized for the temperature 65^oC, operating at supply voltage 0.9V, are shown in Fig. 4. HSPICE simulations for different technology nodes (20nm, 14nm, and 10nm), using BSIM-CMG libraries [24], are used to calculate the delay values at different temperatures, and the temperature profile for different benchmarks has been used to calculate the average performance improvement for each technology node. As a result of the TEI phenomenon in FinFET-based designs, the worst case delay occurs in the lowest temperature. Consequently, delay at the worst case temperature (set to be 25^oC in this analysis) has been chosen as the baseline of this comparison.

Although the buffer insertion technique significantly increases the performance, a considerable power overhead is also imposed to the chip. More precisely, as temperature rises, sub-threshold leakage power increases exponentially. V_{th} degradation in FinFET, exacerbates the situation due to the exponential dependency of leakage power on V_{th} and the quadratic dependency on temperature [7]. For this purpose, an optimal power management mechanism is proposed to minimize the interconnect energy consumption for different temperatures and voltage domains.

B. Energy Minimization Methodology

The repeater power dissipation can be formulated as:

$$P_{repeater} = P_{switching} + P_{short\_circuit} + P_{static}$$

Switching power may be calculated as:

$$P_{switching} = \beta(s(C_g + C_p) + l_{wire} C_{wire}) V_{DD}^2 f_{clk} \quad (13)$$

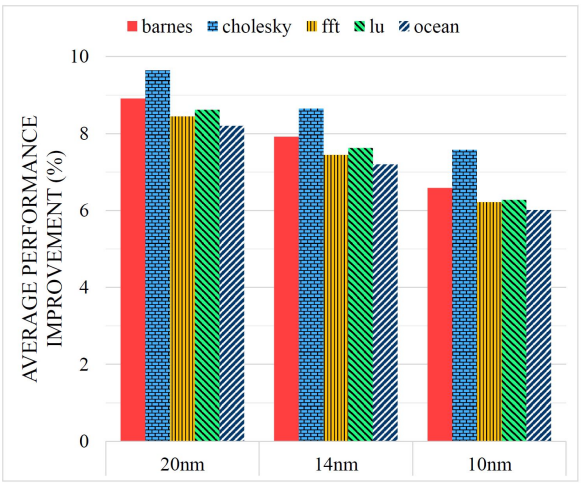

Fig. 4. Average performance improvement (%) of an OLOB interconnect segment of length 1mm optimized for 65^oC at 0.9V in different technologies for various benchmarks compared with worst case delay at 25^oC.

In which f_{clk} denotes the clock frequency and β is the activity factor which captures the probability that a circuit node makes transition from 0 to 1 [20]. Considering sub-threshold current as primary source of static power one could calculate static power as:

$$P_{static} = V_{DD} I_{off} = c_1 V_{DD} \cdot T^2 \cdot (e^{\frac{-V_{th}}{c_2 T}}) \quad (14)$$

where c_1 and c_2 are technology dependent parameters and I_{off} is subthreshold current at $V_{GS} = 0$.

Short-circuit power could be approximately modeled as:

$$P_{short\_circuit} = \frac{1}{12}\beta \cdot ks \cdot V_{DD}(V_{DD} - 2V_{th})^2 \frac{\tau_{in}}{(1 + \frac{\tau_{out}}{\tau_{in}})} f_{clk} \quad (15)$$

where k is a technology dependent parameter and τ_{in} and τ_{out} are input and output transition times. Therefore, total power could be summarized as:

$$P_{total} = c_1 V_{DD} T^2 (e^{\frac{-V_{th}}{c_2 T}}) + c_3 s V_{DD}(V_{DD} - 2V_{th})^2 + (c_4 s + c_5) V_{DD}^2 \quad (16)$$

where c_3 , c_4 and c_5 are calculated based on (13) and (15) and are independent of T, V_{DD} and V_{th}.

The relation between delay per unit length, supply voltage and threshold voltage in buffered interconnects (Eq. 7) could be simplified as:

$$(\frac{\tau}{l})_{opt} \propto \begin{cases} \frac{1}{\sqrt{V_{DD} - V_{th}(T)}} & V_{GS} > V_{th} \\ \sqrt{e^{-\frac{(V_{DD} - V_{th}(T))}{n\nu_T}}} & V_{GS} \leq V_{th} \end{cases} \quad (17)$$

By changing supply voltage or threshold voltage with a negligible performance penalty, we could achieve high energy savings. Hence, our energy minimization methodology considers two optimization knobs (i.e. V_{th} scaling [16] and *dynamic voltage scaling* (DVS)). Threshold voltage scaling of FinFET devices has been proven to be an effective method for leakage power reduction [16]. However, due to the nonlinear relationship between delay, power and temperature, decreasing V_{DD} may lead to higher energy savings in some temperatures.

Algorithm 1 Energy Minimization Methodology in FinFET-based buffered global interconnects

1: Initialize the possible V_{DD} array $\mathcal{V} = [V_{DD}^L, ..., V_{DD}^H]$
2: Measure the temperature T
3: Initialize the possible V_{th} array $\mathcal{B} = [V_{th}^L, ..., V_{th}^H]$ based on current T and possible $\mathcal{V}$ values
4: Compute maximum allowable delay slack σ based on current T and user-defined maximum delay bound
5: Find all possible $(\overline{V_{DD}}, \overline{V_{th}})$ such that $Delay(\overline{V_{DD}}, \overline{V_{th}}, T) \leq (1+\sigma) \cdot Delay(V_{DD}^{nom}, V_{th}^{nom}, T)$
6: Among all possible $(\overline{V_{DD}}, \overline{V_{th}})$ pairs from step 5, find $(V_{DD}^{ME}, V_{th}^{ME})$ with minimum energy

Our energy minimization algorithm is shown in Algorithm 1. At design time, based on different temperatures and user-defined allowable delay bound, the maximum allowable delay slack for each case is determined based on the minimum clock frequency and the estimated performance improvement due to TEI, for all possible (V_{DD}, V_{th}) pairs based on the available voltage steps. Then, based on the total power, Eq. (16), and the propagation delay, Eq. (5), the energy for each possible (V_{DD}, V_{th}) pair is calculated as the power delay product. The results of the delay, power and energy consumption for different values of source voltage, threshold voltage and temperature are stored in lookup tables. Delay values of all the possible (V_{DD}, V_{th}) pairs at different temperatures are stored in a sorted array.

Next, at each decision epoch, based on the current temperature and clock frequency among all the (V_{DD}, V_{th}) pairs which satisfy performance requirements (first k elements of the array), the (V_{DD}, V_{th}) pair with the lowest energy point, namely $(V_{DD}^{ME}, V_{th}^{ME})$, is chosen. The DVS unit takes care of V_{th} and V_{DD} scalings in each decision epoch. Since the temperature does not change abruptly, the epoch should be long enough to capture temperature variations.

V. Simulation Results

We have studied the impact of our proposed energy minimization method on a long (1mm) OLOB interconnect optimized at 65^oC. Delay and power consumption of the interconnect are calculated based on HSPICE simulations using the BSIM-CMG model [24] for a 14nm FinFET technology. Results of interconnect energy savings for zero performance penalty at different temperatures have been reported in Fig. 5. Reducing V_{DD} is mostly effective in lower temperatures where dynamic power dominates the leakage power. On the other hand, at higher temperatures, in which leakage power increases exponentially, threshold voltage scaling is the most promising solution.

It should be noted that performance at the lowest temperature (i.e., 25^oC in this paper) has been chosen to set the target clock frequency. For higher temperatures, based on the available slack, V_{DD} and V_{th} values have been chosen such that the energy consumption is minimized. If the performance constraint is selected for a higher die temperature, more emphasis should be placed on scaling the V_{th} than V_{DD}. To further evaluate our proposed method, we have used the acquired temperature profile of simulation of five benchmarks from SPLASH2 benchmark suite [10] on an 8-core MPSoC using SNIPER [26], to calculate the average interconnect energy saving for each benchmark. The average energy saving for each benchmark is reported in Table I, which shows on

average 16% interconnect energy saving with no performance penalty.

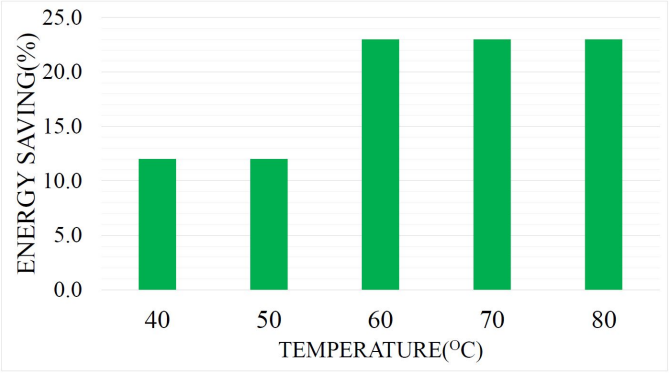

Fig. 5. Energy savings (%) for proposed methodology at each temperature (e.g., at 60°C, interconnect energy could be reduced by 23%) .

TABLE I. Energy savings gained by running Algorithm 1 on five SPLASH2 benchmarks.

Benchmark	Energy Saving	Benchmark	Energy Saving
barnes	17%	cholesky	22%
fft	15%	lu	15%
ocean	13%	**Average**	**16%**

VI. Conclusions

The overall performance of state-of-the-art designs is largely dominated by the delay of global interconnects. FinFET devices have been recognized as a promising replacement for MOSFETs due to superior attributes, such as lower variations and damped short channel effects. One of the most interesting characteristics of such devices, called Temperature Effect Inversion (TEI) phenomenon, leads to lower gate delay in higher temperatures. In this work, we have studied gate delay vs. temperature characteristics of FinFET devices and global interconnects, including the impact of TEI and designed a temperature aware buffer insertion technique. Furthermore, due to the lower delay of an optimum length and optimally buffered (OLOB) interconnect at higher temperatures, we take advantage of V_{DD} and V_{th} scaling techniques to reduce the power consumption with no performance penalty. Our method is not limited to FinFET devices, but is also applicable to conventional long-channel CMOS designs in sub/near-threshold regimes. Under such scenarios, the TEI phenomenon may also be used to enhance the performance or reduce the energy dissipation of global interconnects.

Acknowledgment

This work is supported by the software-hardware foundations program of the CISE Directorate of the National Science Foundation.

References

[1] X. Wang , A. Brown, B. Cheng, and A. Asenov, "Statistical Variability and Reliability in Nanoscale FinFETs," in *IEEE International Electron Devices Meeting (IEDM)*, Dec 2011.
[2] T. Matsukawa *et al.*, "Comprehensive Analysis of Variability Sources of FinFET Characteristics," in *Symposium on VLSI Technology*, 2009.
[3] S. Tang *et al.*, "Finfet - a quasi-planar double-gate mosfet," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2001.
[4] X. Lin, Y. Wang, and M. Pedram, "Joint sizing and adaptive independent gate control for finfet circuits operating in multiple voltage regimes using the logical effort method," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov 2013.
[5] E. Nowak *et al.*, "Turning silicon on its edge [double gate cmos/finfet technology]," *IEEE Circuits and Devices Magazine*, vol. 20, no. 1, 2004.
[6] A. Iranfar, S. Nazar Shahsavani, M. Kamal, and A. Afzali-Kusha, "A heuristic machine learning-based algorithm for power and thermal management of heterogeneous MPSoCs," in *ISLPED*, July 2015.
[7] W. Lee *et al.*, "Dynamic thermal management for finfet-based circuits exploiting the temperature effect inversion phenomenon," in *International Symposium on Low Power Electronics and Design (ISLPED)*, August 2014.
[8] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat. "3-d ics: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," vol. 89, no. 5, May 2001.
[9] International technology roadmap for semiconductors.
[10] S. C. Woo *et al.*, "The SPLASH-2 Programs: Characterization and Methodological Considerations," in *International Symposium on Computer Architecture (ISCA)*, 1995.
[11] K. Banerjee and A. Mehrotra, "A Power-Optimal Repeater Insertion Methodology for Global Interconnects in Nanometer Designs," *IEEE Transactions on Electron Devices*, vol. 49, no. 11, Nov 2002.
[12] C.-P. Chen, Y.-W. Chang, and D. F. Wong, "Fast Performance-Driven Optimization for Buffered Clock TreesBased on Lagrangian Relalxation," in *Design Automation Conference (DAC)*, 1996.
[13] J. Lilis, C.-K. Cheng, and T.-T. Y. Lin, "Optimal wire sizing and buffer insertion for low power and a generalized delay model," vol. 31, no. 3.March 1996.
[14] V. Adler and E. G. Friedman, "Repeater design to reduce delay and power in resistive interconnect," vol. 45, no. 5, May 1998.
[15] A. Nalamalpu and W. Burleson, "A practical approach to DSM repeater insertion: Satisfying delay constraints while minimizing area and power," in *International ASIC/SOC Conference*, 2001.
[16] A. Muttreja, P. Mishra, and N. K. Jha, "Threshold voltage control through multiple supply voltages for power-efficient finfet interconnects," in *International Conference on VLSI Design*, 2008.
[17] T. Sakurai and A. R. Newton, "Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, Apr 1990.
[18] S.-Y. Kim *et al.*, "Temperature dependence of substrate and drain-currents in bulk finfets," vol. 54, no. 5, May 2007.
[19] X. Lin, Y. Wang, and M. Pedram, "Joint sizing and adaptive independent gate control for finfet circuits operating in multiple voltage regimes using the logical effort method," in *2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov 2013.
[20] N. H. E. Weste and D. M. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, 4th ed. San fransisco, CA: Addison-Wesley, 2005.
[21] Y. Pu *et al.*, "Misleading Energy and Performance Claims in Sub/Near Threshold Digital Systems," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov 2010.
[22] A. Calimera, R. I. Bahar, E. Macii, M. Poncino, "Reducing leakage power by accounting for temperature inversion dependence in dual-Vt synthesized circuits," in *International Symposium on Low Power Electronics and Design (ISLPED)*, Aug 2008.
[23] J.-M. Sallesea *et al.*, "A design oriented charge-based current model for symmetric dg mosfet and its correlation with the ekv formalism," vol. 49, no. 3, March 2005.
[24] V. Sriramkumar *et al.* BSIM-CMG 107.0.0: Multi-Gate MOSFET Compact Model (Technical Manual).
[25] S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd ed. Mc-Graw Hill, 2006.
[26] T. Carlson, W. Heirman, and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation," in *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, Nov 2011.
[27] W. Huang *et al.*, "Accurate, pre-rtl temperature-aware design using a parameterized, geometric thermal model," *IEEE Transactions on Computers*, vol. 57, no. 9, Sept 2008.
[28] G. G. Faust *et al.*, "Archfp: Rapid prototyping of pre-rtl floorplans," in *2012 IEEE/IFIP 20th International Conference on VLSI and System-on-Chip (VLSI-SoC)*, Oct 2012.