Electrothermal Engineering in the Nanometer Era: From Devices and Interconnects to Circuits and Systems

Kaustav Banerjee, Sheng-Chih Lin and Navin Srivastava
Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106
{kaustav, sclin, navins}@ece.ucsb.edu

Abstract—*Management of electrothermal (ET) issues arising due to power dissipation both at the micro- and macro- scale is central to the development of future generation microprocessors, integrated networks, and other highly integrated circuits and systems. This paper will provide a broad overview of various ET effects in nanoscale VLSI and highlight both technology and design choices that are thermally-aware. First, effects at the micro scale--in interconnects and devices and their implications for performance, reliability and design are discussed. Next, macro scale--circuit and system level issues including substrate temperature gradients as well as strong ET couplings between supply voltage, frequency, power dissipation and junction temperature in leakage dominant technologies are outlined. A recently developed system level ET analysis methodology and tool that comprehends ET couplings in a self-consistent manner and can generate accurate thermal profile of the substrate is summarized. The application of the ET-tool is demonstrated in a number of areas from power-performance-cooling cost tradeoff analysis to circuit optimization, full-chip leakage estimation, and temperature/reliability aware design space generation. Implications of chip cooling for nanometer scale bulk and SOI based CMOS technologies are also discussed. The ET analysis tool is also shown to be useful for hot-spot management. The paper ends with a brief discussion of electrothermal issues in emerging 3-D ICs and highlights the advantages of employing hybrid Carbon Nanotube-Cu interconnects in both 2-D and 3-D designs.*

I. Introduction

During the past two decades CMOS integrated circuits (ICs) have witnessed unprecedented improvements in their functionality and performance. This was primarily achieved by aggressive technology scaling, which resulted in device density and performance doubling roughly every 18 months as per Moore's law while achieving a remarkable 25% per year improvement in cost per chip function. As CMOS scaled from generation to generation, power dissipation increased proportionately to increasing transistor density and switching speeds. However, with the minimum feature size of the transistor entering the nanometer regime (< 100 nm), power dissipation is increasing ominously especially due to a substantial increase in the leakage power. Leakage power used to be insignificant for earlier generations of ICs but is becoming an increasing fraction of the total power [1]. Moreover, most leakage mechanisms are strongly temperature dependent. This strong coupling between temperature and leakage can cause further increase in total power dissipation. In fact, the International Roadmap for Semiconductors (ITRS) [2] forecasts that high-performance ICs will dissipate around 200 W within a few years. Hence, in spite of the ongoing efforts to reduce power through clever design and technological innovations in the form of new materials and device structures, management of thermal issues will be central to the development of nanoscale VLSI and future generation microprocessors, integrated networks, and other highly integrated systems [1, 3-4]. Since power consumed by the ICs is converted into heat, the corresponding heat densities are rising exponentially. These *electrothermal effects* within the chip are leading to issues and challenges in the design and analysis of high-performance ICs that previous generations did not exhibit [1, 3, 5-7]. For example, since power dissipation in high-performance ICs is spatially non-uniform across the chip and localized heating can occur much faster than chip-scale heating, *hot-spots* and *temperature gradients* are formed that can cause timing errors and reliability problems [8-10]. In short, problems over the heat generated by semiconductor chips are becoming so severe that they threaten to slow, or even limit the development of the entire IC industry. Since, higher temperature slows down the speeds of transistors and interconnects, it is no longer prudent to target highest packing densities in the designs and instead, if the design can be made in a *thermally aware* manner, higher performance and reliability can be achieved for any given technology generation.

In general, these electrothermal effects may impact device, circuit, and system level metrics that are strongly interlinked with one another. For instance, in addition to increasing the leakage, temperature is also known to degrade *device* (due to reduced mobility of electrons and holes) and *interconnect* delay (due to increased resistivity of metal wires arising from increased scattering of electrons with the lattice) and also degrade their reliability (since, mean time-to-failure for reliability mechanisms, including electromigration (EM) and time-dependent gate-oxide breakdown (TDDB), have an inverse exponential dependence on temperature). This in turn leads to degradation of *circuit* level parameters including timing, which eventually degrades *system* level performance. Moreover, several circuit architectures such as dual-V_{th} / V_{dd} designs, and physical design issues including P/G integrity and placement and routing are strongly affected by temperature [7, 11-12]. Furthermore, increased device leakage eventually manifests itself as increased chip power and junction temperature, which in turn, will affect *system level* thermal management (packaging and cooling) solutions and eventual cost per chip function [3, 6]. Therefore, the design process must comprehend thermal effects to ensure improved circuit performance and reliability. Also, in order to formulate comprehensive solutions to the thermal problems, it is necessary to comprehend the correlation between the problems at various levels of the design process, the reliability and performance issues at all levels, and the implications arising due to technology and process conditions. Hence, an integrated holistic approach towards understanding these electrothermal effects at various levels is absolutely vital. **Fig. 1** illustrates the *Electrothermal Engineering* vision to convey its broad scope. Electrothermal Engineering involves development of methods and tools to facilitate the incorporation of temperature and temperature-dependent issues (including reliability) at various stages in the VLSI design process. It also includes exploration of novel thermal management techniques, either through new materials or *thermally-aware* device/interconnect/circuit/system design. We feel that such efforts would be necessary for correctly designing future generations of integrated circuits in a cost effective manner. Furthermore, the lessons learned under electrothermal engineering of nanoscale CMOS based designs are directly applicable to most emerging technologies such as 3-D ICs due to the common issues arising from increased power densities and/or increased temperature sensitivity of performance metrics.

0-7803-9451-8/06/$20.00 ©2006 IEEE.

The paper has been organized as follows. Section II discusses various ET issues at the micro-scale: in devices and interconnects. Section III illustrates macro-scale ET issues in circuits and systems. Electrothermal issues in emerging technologies such as 3-D ICs are discussed briefly in Section IV and the positive implications of employing hybrid carbon nanotube-Cu interconnects in both 2-D and 3-D ICs are highlighted. At this point, it is instructive to rigorously define micro- and macro-scale electrothermal effects. As shown in **Fig. 2**, micro- and macro- scale ET effects differ mainly in the length and time scales involved. Micro-scale ET effects involve length/time scale of 10nm-10μm/0.1ns-10μs, while macro-scale issues involve much larger length/time scales.

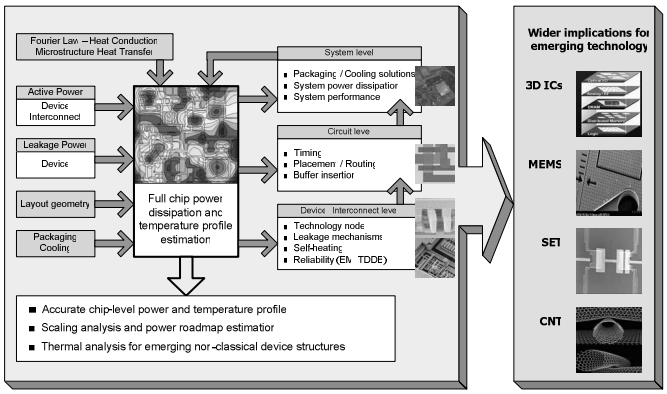

Fig. 1: Schematic overview of the scope of electrothermal engineering research in nanometer scale VLSI and wider implications for emerging technologies such as 3-dimensional ICs, micro-electromechanical systems (MEMS), single electron transistors (SET) and carbon nanotubes (CNT).

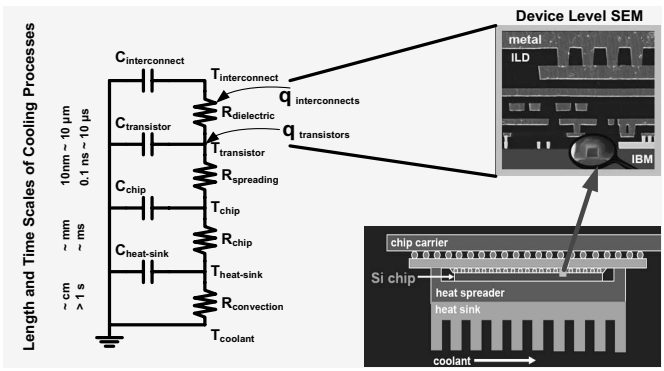

Fig. 2: Global view of heat transfer in integrated circuits: illustrating the different length and time scales of cooling processes at the micro-scale: devices and interconnects and the macro-scale: circuits and system.

II. Electrothermal Engineering at the Micro-Scale: Transistors and Interconnects

II.A. Electrothermal Engineering of Nanoscale Transistors

Conventionally, thermal transport in semiconductors has been modeled through classical diffusion theory [13-14]. However, as device dimensions scale to the orders of tens of nanometer or comparable to the mean free path of phonons (~300nm for bulk Si at 300K), the physics of thermal transport gets affected resulting in localization of heat. At these length scales, the transfer of energy from the charge carriers to the lattice occurs at a faster rate than the relaxation of thermal energy in the lattice, leading to local non-equilibrium in the lattice. Consequently the local hot spot temperature rises beyond the diffusion theory predictions. This is predicted to be a potential obstacle to the continued scaling of devices. This effect has been illustrated through simulations involving the solution of the Boltzmann Transport Equation (BTE) in MOSFETs (**Fig. 3**) [15]. This also has significant implications for device reliability. For example, it has been shown that under electrostatic discharge (ESD) conditions, which involves very high current (> few Amps) for a very short time (<100 ns), the classical diffusion theory will not be able to predict the failure temperature accurately leading to poor correlation between measurements and simulations [16]. This effect is expected to become even more important for sub-90 nm technologies, where it can influence other parameters such as device delay and leakage.

Furthermore, it has been predicted that scaling CMOS to and beyond the 22-nm technology node will probably require the introduction of several new material and structural changes to the MOSFET to sustain performance increases as per the ITRS projections and to manage short channel effects as shown in **Fig. 4** [17-18]. These new technologies have been shown to alleviate short channel effects, reduce leakage, increase drain saturation current at even lower operating voltage. Material changes will include strained silicon n- and p-channels, high-k gate dielectric and metal gate electrode. Structural changes could include fully depleted Ultra Thin Body (UTB) SOI single- gate MOSFETs, followed by fully depleted UTB double-gate structures. Strained-Si increases drive current due to strain induced enhanced mobility and average carrier velocity. The UTB SOI MOSFET consists of a very thin ($t_{si} \leq 10$ nm), fully depleted (FD) transistor body to ensure good electrostatic control of the channel by the gate in the off state. The use of a lightly doped or undoped body enhances carrier mobility for higher transistor drive current.

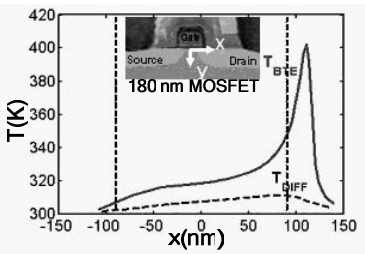

Fig. 3: Temperature distribution along the channel region [15].

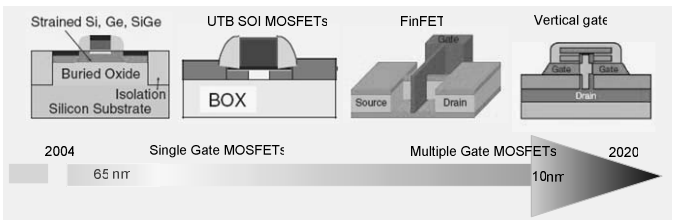

Fig. 4: Evolution of CMOS technology towards transport-enhanced (Strained-Si, SiGe) and UTB devices. These devices exhibit poorer thermal properties due to confined geometry and poor thermal conductivity materials.

However, the use of buried oxide in SOI type devices or SiGe graded layer in Strained-Si devices increases the thermal resistance of the device due to low thermal conductivity of these materials [19, 20]. It has been shown that drain current can reduce by as much as 15% due to self-heating in a 100 nm strained-Si device [19]. Also, the thermal conductivity of thin Si layer degrades due to increased phonon boundary scattering [21], which in turn, results in poor thermal properties of the channel layer in an UTB device [21-23]. Furthermore, in highly confined geometries such as double-gate and tri-gate structures, thermal resistance is further exacerbated due to multiple interfaces. The poor thermal properties of these non-classical devices results in higher temperature rise and subsequently lower drive current and higher leakage than predicted.

Consequently, it is extremely important to analyze and optimize these electrothermal effects in non-classical devices for adequate use of these in nanoscale circuits and systems. For example, ongoing work in our group indicates that in FinFET devices while reducing

the fin thickness gives improved channel control, self-heating increases substantially due to increased phonon-boundary scattering. Hence, there is a trade-off between channel control and junction temperature. Similar electrical-thermal trade off exists for other FinFET device parameters such as fin height, BOX thickness and fin pitch for a multi-fin structure. Thus, the scaling of silicon transistors will, in general, lead to increased temperature rise. However, understanding the physics of heat conduction in the device may provide an additional way of controlling the temperature rise and allow devices to be operated closer to their peak performance levels. Existing diffusion based thermal transport models underestimate the peak temperature rise in these non-classical devices where non-equilibrium thermal transport can lead to dramatically high temperatures.

Micro-scale electro-thermal modeling in the literature can be widely separated into two categories. The first method considers heat generation and transport as two separate problems and solves them rigorously, resulting in non self-consistent and inaccurate results. The second method develops a fully coupled model making simplistic assumptions [24]. However, with device scaling these coupled solutions become more and more inaccurate due to assumptions invalid at nanoscale. We are developing a fully coupled method while removing these assumptions by employing more sophisticated methods for heat generation via electron Monte Carlo simulations and for heat conduction via phonon BTE [25-26]. Once the electron MC and the phonon BTE methods are developed, they will be coupled together for electrothermal simulations as shown in **Fig 5**. Electron scattering mechanisms are temperature dependent, therefore the temperature distributions along the device will be fed back to the MC simulations. The resulting heat generation profile will be fed back into the phonon BTE for self-consistent solutions.

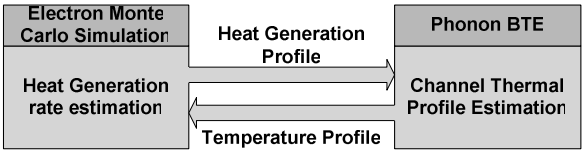

Fig. 5: Schematic illustrating the coupled electron Monte-Carlo and phonon BTE approach.

II.B. Electrothermal Engineering of Nanoscale Interconnects

With the aggressive scaling of VLSI technology, cross-sectional dimensions of on-chip interconnects in current technologies [2] are of the order of the mean free path of electrons in copper (40 nm at room temperature). At such dimensions, the increase in Cu interconnect resistivity due to the presence of a highly resistive barrier layer (which occupies a significant fraction of the drawn wire width) is further exacerbated by the increased scattering of electrons at the surface and grain boundaries, as shown in **Fig. 6(a)** [27]. The low-k dielectric materials used for intra- and inter- metal layer dielectric (ILD) in current VLSI technologies have much lower thermal conductivities than silicon dioxide (**Fig. 6(b)**) [27]. Alongside the lower thermal conductivity of inter-layer dielectrics, the scaling of technology also results in higher current density demands on interconnects [28].

The combined effect of increasing copper interconnect resistivity, decreasing thermal conductivity of ILD materials and rising current densities in on-chip wires results in significant rise in interconnect temperatures, especially at the global metal layers which are furthest away from the heat sink (**Fig. 7(a)**) [27]. These high temperatures will become a major concern for interconnect reliability as the mean time to failure due to electromigration depends exponentially on metal temperature. The maximum current density that can be supported by these interconnects will thus be severely limited due to reliability constraints (**Fig. 7(b)**) [28].

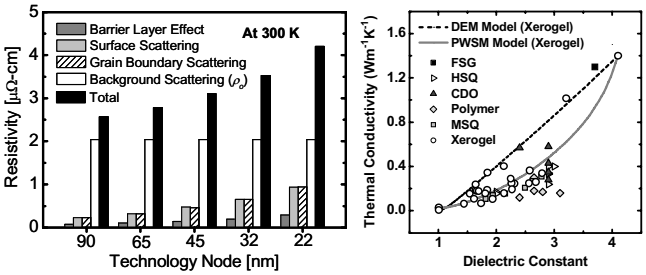

Fig. 6: (a) Scaling of Cu resistivity for the ITRS intermediate wires (at 300K). The total resistivity is the sum of all resistivity components. Parameter values are: ρ_o=2.04 μΩ-cm (300K), λ=37.3 nm (300K), p=0.41, and R=0.22. **(b)** Correlation of the thermal conductivity and dielectric constant showing the simultaneous decrease of thermal conductivity and dielectric constants of several materials along with physical models that predict this behavior [27].

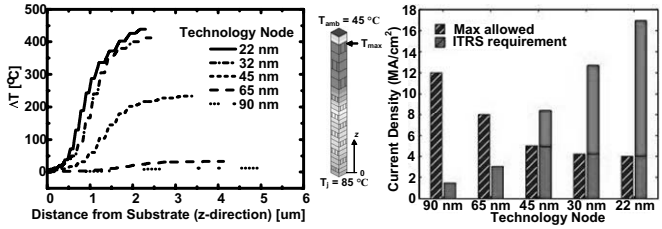

Fig. 7: (a) Temperature rise (ΔT) along the vertical distance from the substrate at different technology nodes [27]. The temperature contour plot of 45 nm technology node is also shown as an example. **(b)** Maximum allowed current density (duty ratio=0.001) in local vias from self-consistent electromigration lifetime estimation vs. the ITRS requirement for current density [28].

III. Electrothermal Engineering at the Macro-Scale: Circuits and Systems

III.A. Circuits

III.A.1. Interconnect Power Dissipation

While typical local interconnect delay is expected to decrease with technology scaling (mainly as a result of the higher packing density of devices), global interconnect delay increases. In order to keep the delay of global interconnects under control, repeaters (buffers) are inserted at regular intervals to drive signals faster. As technology scales, an increasing number of buffers is required for the global interconnection system on a chip (**Fig. 8**). These repeaters can contribute significantly to total chip power dissipation, which is a critical problem for high-performance ICs.

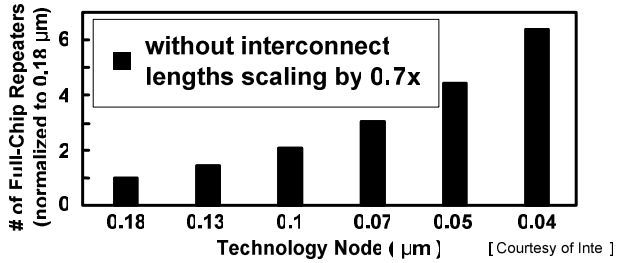

Fig. 8: Number of repeaters increases rapidly with technology scaling.

Increase in effective wire resistivity will further increase wire delay resulting in more numerous repeaters and larger chip power dissipation. However, it has been shown that when delay is not of critical importance "power-optimal" repeater insertion [29] can be used to achieve large power savings. This methodology assumes tremendous importance in the light of power-limited technologies of current and future IC generations as the power savings for a given amount of delay penalty increases as technology scales (**Fig. 9(a)**). This is mainly due to the increasing leakage power dissipation in

CMOS devices (**Fig. 9(b)**). In the presence of variations in devices as well as interconnects, the power optimal repeater insertion can be significantly impacted as shown in **Fig. 10** [30]. Also, as shown in **Fig. 11**, for a given delay penalty, power savings are greater under higher percentage of variations, mainly because of the increase in leakage power under variations [30]. As a result the impact of the power-optimal repeater insertion methodology becomes more significant in aggressively scaled technologies which are more susceptible to variations.

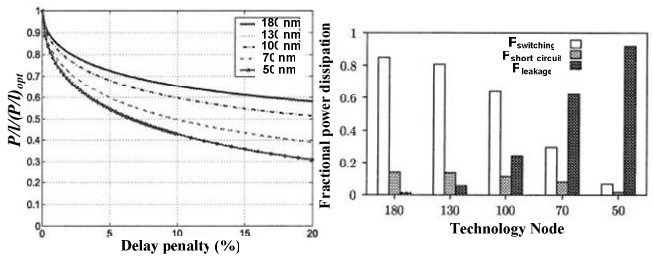

Fig. 9: (a) Normalized power per unit length for a buffered global interconnect (normalized to power per unit length for optimally buffered case) as a function of delay penalty, at different technology nodes. **(b)** Relative contributions of three components of overall power dissipation for 5% delay penalty at different technology nodes [29].

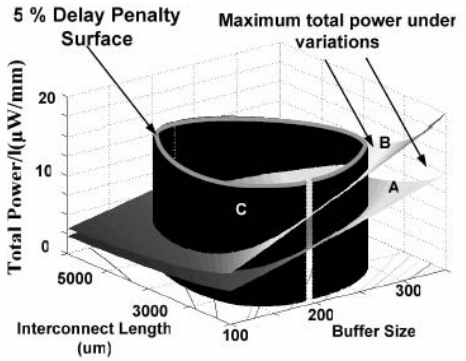

Fig. 10: Maximum value of total power per unit length as a function of buffer size (s) and interconnect length (l) under two different cases (amount of parameter variations is much higher in Case B as compared to Case A) [30]. Also drawn is a delay penalty surface, along which $(\tau/l) = 1.05\ (\tau/l)_{opt}$. The 2-D contour that gives 5% delay penalty has been extended in 3-D plane (surface C) for better visualization.

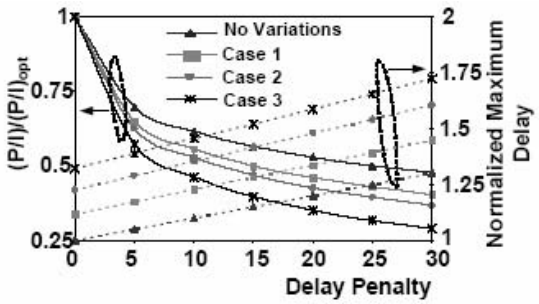

Fig. 11: Power per unit length normalized to optimal power per unit length, for different delay penalties under different conditions [30]. Normalized maximum delay curves are also drawn. Parameter variations increase from Case 1 to Case 3, as 10%, 20% and 30% of mean value.

III.A.2. Impact of On-Chip Thermal Gradients

Recent work indicates that, in high performance ICs, the peak chip temperature can rise up to 140ºC in 90 nm technology node and is expected to rise even further for future technology nodes [31]. Since, these peak temperatures always occur at the top of the chip, they can significantly increase the interconnect resistance, which would in turn increase the signal propagation delay in the interconnect line. In nanometer scale high performance ICs, large temperature gradients can also occur in the substrate. These gradients, for example, may be created due to "spotty" gate-level switching activity and/or because various functional blocks are put in different operational modes, e.g., active, standby, or sleep modes [32]. Dynamic power management (DPM) [33] and clock gating can also be major contributors to a non-uniform substrate temperature. In fact, thermal gradients as large as 50°C can exist across high-performance microprocessor substrates (**Fig. 12**). **Fig. 13** shows the error in interconnect delay that can occur as a result of assuming a uniform average temperature for two different temperature gradients existing along the line. **Fig. 14** depicts the percentage skew between wires 1 and 2 (which are of identical length = 2000 μm) as a function of the position x where the thermal gradient occurs, while wire 1 is at a uniform temperature of $T_1 = 100°C$. **Fig 15** shows that by neglecting the thermal effects of hot spots on the resistivity of the global layers, worst-case voltage drop can not be predicted correctly. Finally, **Fig 16** shows the difference resulting from using a temperature-aware buffer insertion technique and the increasing improvement in performance that can be obtained as the thermal gradient increases.

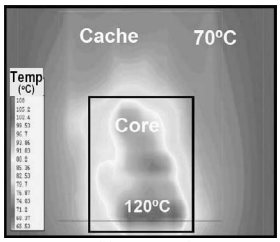

Fig. 12: Varying temperature profile on a microprocessor [courtesy S. Borkar, Intel].

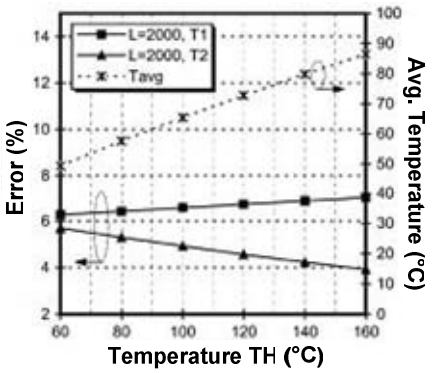

Fig. 13: Percentage delay differences between the non-uniform temperature-dependent interconnect delay and the delay at a uniform line temperature T_{avg}. T_1 and T_2 denote positive and negative exponential temperature gradients between a low temperature T_L of 40°C and a high temperature T_H shown on the x-axis [10].

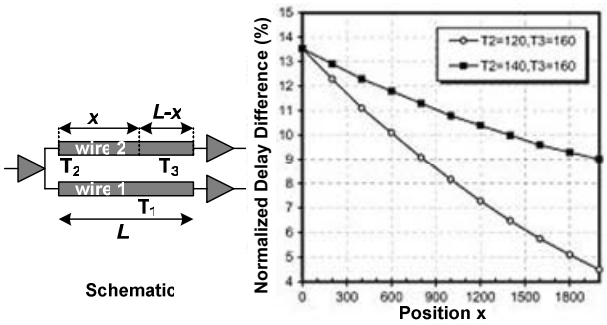

Fig. 14: Percentage of normalized delay difference between wire segment 1 and wire segment 2 as a function of break point *x*, the point at which the thermal gradient occurs in wire segment 2 [10].

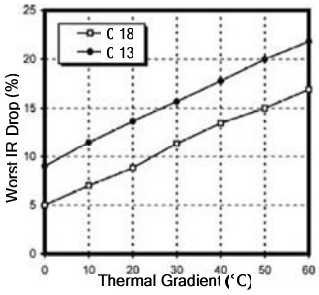

Fig. 15: Worst-case voltage drop (V_{IR}/V_{dd}) increase in presence of hot spots modeled by constant-peak Gaussian distribution as a function of thermal gradient magnitudes (°C), shown for two technology feature sizes [34].

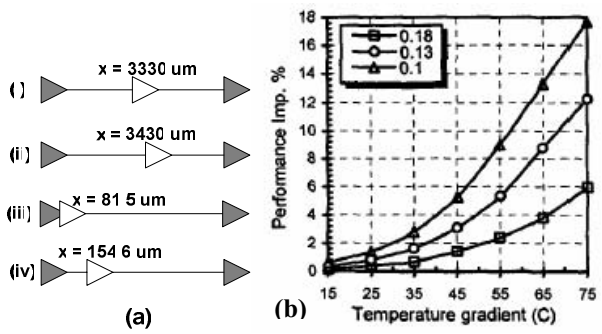

Fig. 16: (a) Location of an inserted buffer in a 6660 um line (180 nm node): (i) standard technique, (ii-iv) temperature aware technique with only variable interconnect resistance (ii), only variable driver resistance (iii) and variable interconnect and driver resistances (iv). **(b)** Delay improvement due to the thermally aware buffer insertion for one buffer as a function of different thermal gradients between the two ends of the line in comparison with the standard buffer insertion techniques for different technologies [35].

III.B. Systems

III.B.1. Electrothermal (ET) Couplings in Nanoscale ICs

Fig. 17(a) illustrates the interdependencies between various design parameters including performance (frequency), power dissipation, supply voltage, threshold voltage, and substrate (junction) temperature [36]. The total power dissipation has two major components: switching and leakage power dissipation. The switching power increases as the chip frequency (performance) and supply voltage increase. Moreover, the performance itself is dependent on temperature due to the dependence of the transistor on-current on substrate temperature. Continuous increase of integration density and power consumption elevates on-chip temperature. Higher power dissipation and temperature of the chip, which further increases the subthreshold leakage, thereby creates a strong feedback loop leading to various ET couplings that used to be inconspicuous for earlier generation of ICs. By taking into account these couplings, a system level ET analysis methodology and tool has been developed [36], using which design tradeoffs between power-performance-cooling cost can be evaluated as shown in **Fig. 17(b)**.

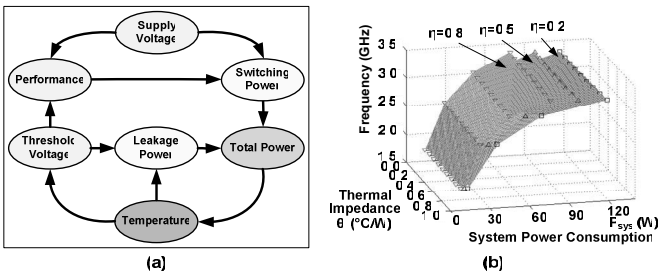

Fig.17: (a) Schematic view of ET couplings between different design parameters. **(b)** Power-performance-cooling cost tradeoff analysis [36].

III.B.2. Implications for Circuit Optimization

For power-constrained applications, lowering supply voltage (V_{dd}) offers the biggest potential to decrease the active power consumption, since CMOS switching power has a quadratic dependence on supply voltage. On the other hand, lowering supply voltage degrades the performance of circuits. It is, however, possible to maintain the performance by decreasing the threshold voltage (V_{th}) at the same time, but then the subthreshold leakage power increases exponentially. Consequently, the need for low power and high performance circuit applications motivates the search for an optimal set of supply and threshold voltages to tradeoff performance and power consumption.

Traditionally, circuit designers use energy-delay product (EDP) as the design metric to minimize both power and delay of a circuit [37]. **Fig. 18(a)** has been generated simply by direct numerical evaluation of energy and delay for a specific design. However, as pointed out in the preceding discussion, it is crucial to incorporate electrothermal couplings when evaluating the power and delay. Recently, an electrothermally coupled EDP methodology has been developed [38]. This methodology takes these electrothermal couplings into consideration and incorporates both analytical models and results from the circuit simulator based on an integrated device, circuit, and system level modeling approach. In comparison with **Fig. 18(a)**, it can be observed from **Fig. 18(b)** that not only the EDP contours and iso-performance curves shift but also the design space gets restricted by thermal constraint that cannot be predicted by traditional evaluation. Consequently, if electrothermal couplings are not considered, power dissipation and delay evaluations will be inaccurate and mislead the design optimization process.

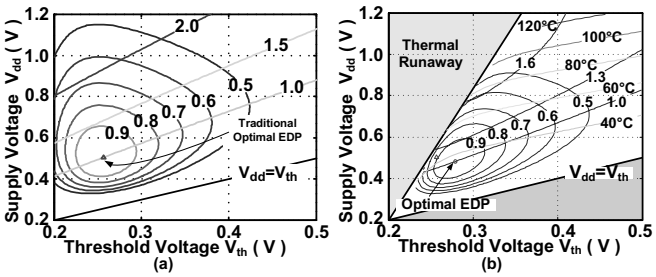

Fig. 18: (a) Traditional EDP evaluation. The EDP contours can be found by normalizing with respect to the value of the EDP at the optimal point as indicated by 'Δ'. For instance, any point on the curve labeled *0.5* has an EDP value twice that of optimal (EDP = $2 \cdot EDP_{opt}$), i.e., minimum value. The numbers on the iso-performance curves indicate the normalized value of the frequency with respect to the frequency at EDP_{opt}. **(b)** Electrothermally coupled EDP evaluation. Note that the design space gets restricted by thermal constraint (thermal runaway) that is determined by a passive cooling model [36], assuming junction-to-ambient thermal resistance $\theta_j = 0.85$ °C/W.

Most reliability mechanisms are highly temperature sensitive. The electrothermally coupled EDP methodology also provides a reliability and thermally aware "design space" that can be used to optimize and compare various designs and also evaluate the relative impact of various reliability constraints on a given design space.

Different reliability and thermal constraints can be applied at the design stage by using the iso-temperature curves shown in **Fig. 18(b)**. It can be observed that the junction temperature must be maintained below 40ºC to achieve optimal EDP and if the junction temperature is required to be 100ºC due to some cost constraints then the EDP will be higher by a factor of 2.5.

In order to highlight the impact of technology scaling on this methodology, **Fig. 19(a)** shows the reduction of design space due to increasing I_{off}. On the other hand, lowering of the junction temperature by employing advanced packaging and cooling techniques with lower thermal impedance (θ_j) will expand the design

space as shown in **Fig. 19(b)**. Thus, tradeoffs between cooling cost and the benefit from design space relaxation also can be evaluated by the proposed methodology. Moreover, the significance of applying this methodology is expected to increase when parameter variations such as process, supply voltage and temperature variations are also taken into account since they are known to increase subthreshold leakage significantly.

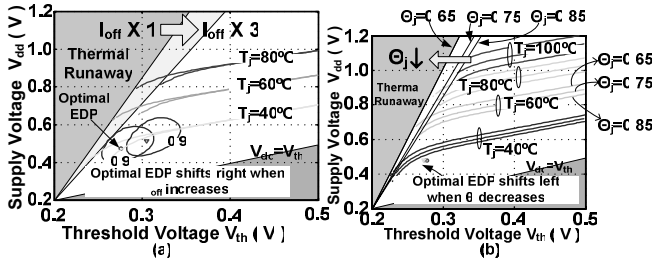

Fig. 19: (a) Impact of technology scaling on design space. While the leakage increases due to technology scaling or process variations, the operation region prohibited by thermal runaway expands. **(b)** Relaxation of design space by lowering of thermal impedance (θ_j). The iso-temperature curves move upwards by employing advanced packaging and cooling techniques with lower θ_j and the operation region prohibited by thermal runaway reduces. Therefore, the design space expands.

Different optimization metrics result in different design choices. Besides energy-delay product (EDP), other design metrics are also used for different applications such as power-delay product (PDP) and power-energy product (PEP). Moreover, in [39], Pénzes and Martin showed that the Et^n metric characterizes any feasible trade-off. Hofstee [40] conclude that optimal metric is not unique for all designs but depends on the desired level of performance.

In order to choose an appropriate design metric in a systematic manner, an electrothermally-aware methodology has been developed [38]. The method captures the relative importance of power dissipation and performance to achieve design-specific targets and also provides a more meaningful basis to optimize supply and threshold voltages as they change from one technology generation to the next. As shown in **Fig. 20**, the optimal operating locus obtained for different optimization metrics based on a generalized metric (PT^{μ}) shifts when technology scales from 100nm to 70nm nodes where μ represents the ratio of exponent of delay to that of power.

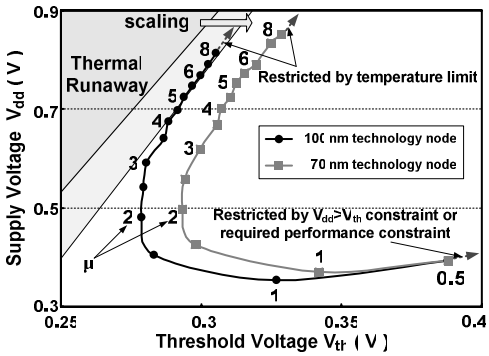

Fig. 20: Electrothermally-aware optimal operation locus for different μ. Note that the region (thermal runaway) expands due to technology scaling.

III.B.3. Implications for Full-Chip Subthreshold Leakage Estimation

For nanometer scale CMOS technologies, leakage power forms a significant component of the total power dissipation, especially due to within-die and die-to-die variations in process (P), temperature (T) and supply voltage (V). As a result of these variations, leakage power has been reported to have 20X variations for a 180 nm CMOS technology [41]. Thus, designing with the worst-case leakage values may result in excessive guard-banding while underestimating the leakage might result in highly optimistic designs. Additionally, due to a 5X increase of total leakage power every generation [42], the design constraint based on leakage power may soon limit the yield. Under this scenario, a probabilistic framework for accurately estimating full-chip subthreshold leakage power distribution under P-T-V variations has been developed, which can be subsequently used to accurately estimate the yield [43].

Fig. 21 shows that apart from within-die variations, die-to-die parameter variations such as channel length and temperature can strongly impact the leakage power. It can be clearly observed from the figure that die-to-die temperature variations significantly increase the leakage due to the electrothermal couplings between subthreshold leakage power dissipation and die temperature, especially at higher operating temperatures.

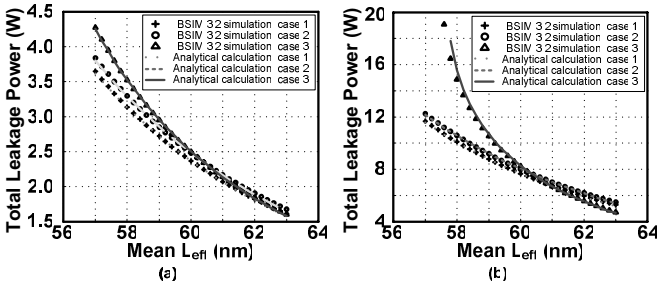

Fig. 21: Total subthreshold leakage power vs. mean die-channel length. **(a)** Average junction temperature = 300K **(b)** Average junction temperature = 320K. In case 1, only die-to-die channel length variations are considered. In case 2, besides die-to-die channel length variations, within-die channel length variations are also considered, while case 3 considers die-to-die temperature variations together with all the variations considered in case 2.

III.B.4. Implications for IC Cooling and Hot-Spot Management

Cooled chip operation is being seriously evaluated as a practical technique for boosting the performance of high-end microprocessors. While the drive current increases at lower temperature, it is observed that SOI type transistors show greater sensitivity to temperature due to the less increase of body to source voltage (V_{BS}) of PD-SOI at low temperature [44] which causes a smaller increase in the saturated threshold voltage of PD-SOI type transistors, as shown in **Fig. 22(a)** [45]. Thus, the enhancement of drive current of PD-SOI transistors at low temperature is higher than that in bulk transistors.

It has been shown that cooling leads to benefit at the device and circuit levels [45]. However, in order to understand the real benefit from cooling, system level considerations need to be taken into account. **Fig. 22(b)** shows the leakage power dissipation for two identical test microprocessor designs at different technology nodes under the application of active cooling. As expected, the leakage power decreases at lower temperature and the reduction of leakage becomes greater as technology scales. **Fig. 23** shows that the chip power (including active and leakage power dissipation) decreases as more cooling is applied mainly as a result of decreasing leakage power which is shown in **Fig. 22(b)**. However, the total system power (including chip power and cooling power consumption), determines the practical limit beyond which further cooling does not lead to any overall power savings and the limit occurs at an increasingly lower temperature as technology scales. Thus, as technology scales, the benefit that can be derived from cooling increases.

Highly integrated circuits with different functional blocks and finite thermal conductivity of silicon and packaging materials will create a non-uniform power density across the chip surface. Those

regions with higher heat flux densities (hot-spots) affect performance and reliability and lead to a general over-design in the microprocessor packaging and cooling solutions. In order to comprehend the impact of cooling on thermal gradients and hot-spots, a self-consistent electrothermal methodology for estimating substrate temperature profile has been developed (**Fig. 24**) [45].

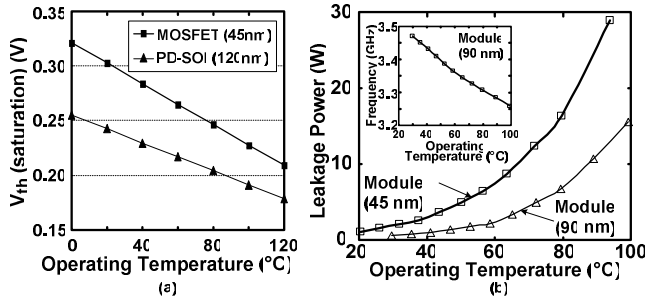

Fig. 22: (a) The increasing rate of saturated threshold voltage for bulk MOSFET (0.9mV/°C) is larger than PD-SOI type transistor (0.6mV/°C). **(b)** Leakage power dissipation as a function of operating temperature. The inset shows chip performance increases as operating temperature decreases [45].

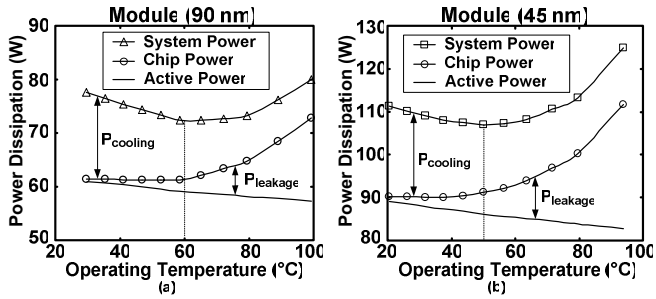

Fig. 23: Electrothermally-aware system level evaluation of power dissipation [45]. A minimum system power determines the practical limit beyond which further cooling does not lead to any power saving. **(a)** 90 nm test module **(b)** 45 nm test module.

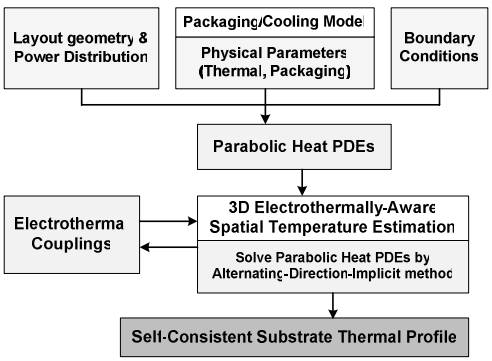

Fig. 24: Overview of the electrothermally-aware substrate thermal profile generator [45].

An example chip design (die size: 10 mm × 10 mm) with power densities per functional block is shown in **Fig. 25(a)**. The spatial substrate temperature profile, **Fig. 25(b)**, shows several hot-spots and the highest junction temperature is around 133°C. Although the results shown here are specific to the above mentioned IC, the conclusions drawn are more generic. **Fig. 26** shows the effect of applying global and localized cooling strategies on hot-spot management. As shown in **Fig. 26(a)**, a lower junction-to-ambient thermal resistance (θ_j) reduces the maximum junction temperature by applying global cooling (through better interface material, higher cooling efficiency, etc.). However, on-chip hot-spots and thermal gradients still remain. On the other hand, localized cooling solutions such as local spray cooling, thin-film thermoelectric coolers, can be applied to electronic application to eliminate the hot-spots. For example, if two thin-film thermoelectric coolers can be placed on the backside of the wafer below the locations of hot-spots, as shown in **Fig. 26(b)**, it can effectively eliminate the targeted hot-spots.

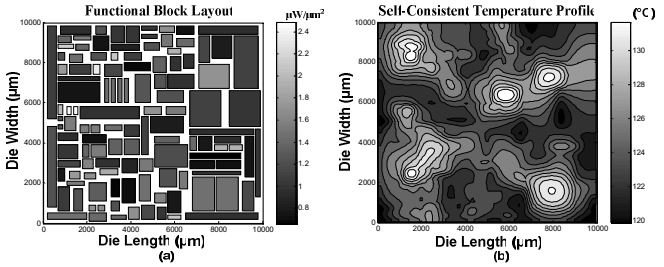

Fig. 25: (a) Functional block layout of a test chip showing power density associated with each block. Nominal total power consumption is 75 W. **(b)** Spatial substrate temperature profile generated using methodology shown in **Fig. 24** (θ_j=1.1 °C/W). Five hot-spots can be observed [45].

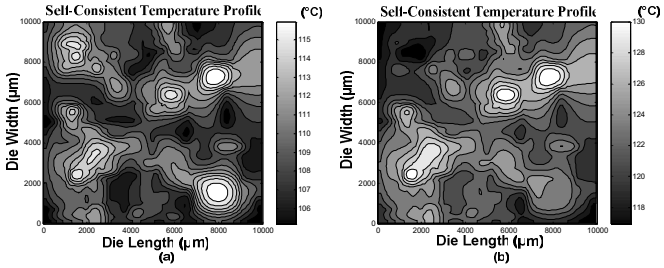

Fig. 26: Spatial substrate temperature profile: **(a)** Applying global cooling (θ_j reduces 20%). Although highest temperature decreases, the hot-spots remain. **(b)** Integrating two thin-film thermoelectric coolers at the top-left and bottom-right hot-spots. Only three hotspots can be observed [45].

IV. Electrothermal Engineering in Emerging Technologies

IV.A. Three-Dimensional (3-D) ICs

The challenges from on-chip interconnects in nanometer scale VLSI have led researchers to seek innovative design solutions, circuit or interconnect optimization techniques and material solutions so that the chip's wires do not offset the benefits of continued device scaling. Three-dimensional integration to create multi-layer ICs [46] is a concept that reduces the number and the average lengths of the longest global wires seen in traditional 2-D chips by providing shorter "vertical" paths for connection and also allows integration of disparate technologies and substrates. However, this technology still needs to overcome difficult challenges such as the development of new system architectures and tools. A critical problem in 3-D ICs is the thermal management of internal (stacked) active layers [31] (**Fig. 27(a)**). Hence, the electrothermal couplings described in Section III.B.1 also have a significant impact on the temperature and power dissipation of 3-D ICs (**Fig. 27(b)**).

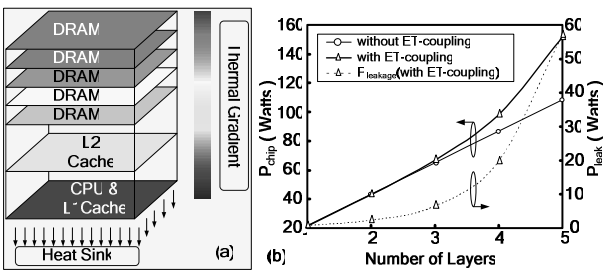

Fig. 27: (a) The prevalent high temperatures on stacked layers in a 3-D integrated circuit. Performance evaluation of such chips must account for the negative impact of these high temperatures. **(b)** Chip power and leakage power dissipation vs. number of layers. As the number of layers is increased the total power dissipation becomes a strict nonlinear function of the number of layers as opposed to a linear function observed by non-self-consistent treatment of the problem.

IV.B. Hybrid Carbon Nanotube-Cu Technologies

Carbon nanotube (CNT) bundle interconnects are possibly the least disruptive of all alternatives to copper interconnects that have been suggested so far. Although there are some technological issues that must be resolved before CNT interconnects can be used in practice [28], they have the potential to meet interconnect challenges without the need for paradigm changes in VLSI circuit design techniques and tools or extra circuitry. Besides the performance benefits of CNT bundle interconnects [47], their high thermal conductivity makes them very effective in controlling the large backend temperature rise expected with metallic interconnects (**Fig. 28**), and hence in improving overall interconnect performance and lifetime [48]. The thermal advantage of CNT bundle vias will also have significant implications for 3-D ICs where thermal management is a big concern.

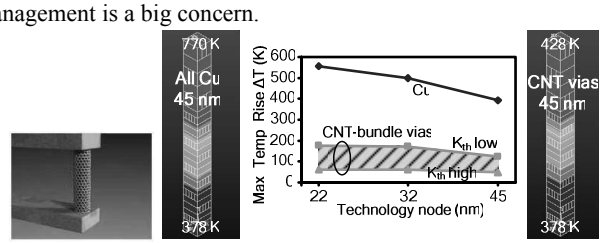

Fig. 28: Maximum interconnect temperature rise for Cu interconnect stack with Cu vias compared to CNT bundle vias integrated with Cu interconnects (see schematic on left [courtesy Infineon]). For CNT bundles, the shaded region shows the range of thermal conductivity 1750 W/mK < Kth < 5800 W/mK [48]. Reference (substrate) temperature = 378K.

Acknowledgements

This work was supported by Intel Corp., Fujitsu Labs. of America, Mentor Graphics Corp., NIST, SRC and the University of California-MICRO program.

References

[1] V. De and S. Borkar, "Technology and design challenges for low power and high performance microprocessors," *ISLPED*, 1999, pp. 163 –168.
[2] International Technology Roadmap for Semiconductors (ITRS), 2004
[3] R. Viswanath et al., "Thermal performance challenges from silicon to systems," *Intel Technology Journal 3rd quarter*, 2000.
[4] I. Aller et al., "CMOS circuit technology for sub-ambient temperature operation," *ISSCC*, 2000, pp. 214-215.
[5] S. Borkar et al., "Parameter variations and impact on circuits and microarchitecture," *DAC*, 2003, pp. 338-342.
[6] P. E. Ross, "Beat the heat," *Spectrum, IEEE*, Vol. 41, pp. 38-43, 2004.
[7] C. X. Zhang, "Timing-, heat- and area-driven placement using self-organizing semantic maps," *ISCAS*, 1993, pp. 2067-2070.
[8] Y-K Cheng et al., "ILLIADS-T: An electrothermal timing simulator for temperature-sensitive reliability diagnosis of CMOS VLSI chips," *IEEE TCAD*, Vol. 17, pp.668-681, 1998.
[9] C. C. Teng et al., "iTEM: A Temperature-dependent electromigration reliability diagnosis tool," *IEEE TCAD*, Vol. 16, pp. 882-893, 1997.
[10] A. H. Ajami, K. Banerjee and M. Pedram, "Modeling and analysis of non-uniform substrate temperature effects on global ULSI interconnects", *IEEE TCAD*, Vol. 24, pp. 849-861, 2005.
[11] J. Lee, "Thermal placement algorithm based on heat conduction analogy," *IEEE Trans. on Components and Packaging Tech.*, Vol. 26, pp.473-482, 2003.
[12] C. H. Tsai and S. M. Kang, "Cell-level placement for improving substrate thermal distribution," *IEEE TCAD*, Vol. 19, pp. 253-266, 2000.
[13] L.T. Su, et al., "Measurement and modeling of self-heating in SOI nMOSFETs", *IEEE TED*, Vol. 41, pp. 69-75, 1994.
[14] Y-K. Leung, et al., "Heating mechanisms of LDMOS and LIGBT in ultra-thin SOI", *IEEE EDL*, Vol. 18, pp. 414-416, 1997.
[15] E. Pop et al., "Localized heating effects and scaling of sub-0.18 micron CMOS devices," *IEDM*, 2001, pp. 677-680.
[16] P. G. Sverdrup et al., "Sub-continuum thermal simulations of deep sub-micron devices under ESD conditions," *SISPAD*, 2000, pp. 54-57.
[17] A. M. Ionescu and K. Banerjee, *Emerging Nanoelectronics: Life With and After CMOS*, Springer, 2005.
[18] T. Skotnicki et al., "End of CMOS scaling", *IEEE Circuits and Devices Magazine*, Vol. 21, pp. 16-26, 2005.
[19] K. A. Jenkins and K. Rim, "Measurement of the effect of self-heating in strained-silicon MOSFETS", *IEEE EDL*, Vol. 23, pp. 360-362, 2002.
[20] W. Liu and M. Asheghi, "Thermal modeling of self-heating in strained-silicon MOSFETs", *IEEE Inter society conference on thermal phenomena*, Vol. 2, pp. 605-609, 2004.
[21] W. Liu and M. Asheghi, "Thermal conductivity of ultra-thin single crystal silicon layers, part I - experimental measurements at room and cryogenic temperatures," *J. Heat Transfer*, 2004.
[22] E. Pop et al, "Thermal analysis of ultra-thin body device scaling", *IEDM*, 2003, pp. 883-886.
[23] M. Asheghi et al., "Thermal conduction in doped single-crystal silicon films," *JAP*, Vol. 91, pp. 5079-5088, 2002.
[24] J. Lai and A. Majumdar, "Concurrent thermal and electrical modeling of sub-micrometer silicon devices," *JAP*, Vol. 79, pp. 7353–7363, 1996.
[25] E. Pop et al., "Analytic band Monte Carlo model for electron transport in Si including acoustic and optical phonon dispersion," *JAP*, 96, 4998, 2004.
[26] S Sinha et al., "A split-flux model for phonon transport near hotspots", *J. Heat Transfer*, 2005.
[27] S. Im et al., "Scaling analysis of multilevel interconnect temperatures for high performance ICs", *IEEE TED,* 2005 (in press).
[28] N. Srivastava and K. Banerjee, "A comparative scaling analysis of metallic and carbon nanotube interconnections for nanometer scale VLSI technologies," *VMIC*, 2004, pp. 393-398.
[29] K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE TED*, Vol. 49, pp. 2001-2007, 2002.
[30] V. Wason and K. Banerjee, "A probabilistic framework for power-optimal repeater insertion for global interconnects under parameter variations," *ISLPED*, 2005, pp. 131-136.
[31] S. Im and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs," *IEDM,* 2000, pp. 727-730.
[32] Z. Yu et al., "Full chip thermal simulation," *ISQED,* 2000, pp.145-149.
[33] Q. Wu, Q. Qiu, and M. Pedram, "Dynamic power management of complex systems using generalized stochastic Petri nets," *DAC*, 2000, pp. 352-356.
[34] A. H. Ajami, K. Banerjee and M. Pedram, "Scaling analysis of on-chip power grid voltage variations in nanometer scale ULSI", *J. of Analog Integrated Circuits and Signal Processing*, Vol. 42, pp. 277-290, 2005.
[35] A. H. Ajami, K. Banerjee and M. Pedram, "Analysis of substrate thermal gradient effects on optimal buffer insertion", *ICCAD*, 2001, pp. 44-48.
[36] K. Banerjee et al., "A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management," *IEDM*, 2003, pp. 893-896.
[37] M. Horowitz et al., "Low power digital design," *ISLPED*, 1994, pp. 8-11.
[38] S-C. Lin et al., "A thermally aware methodology for design-specific optimization of supply and threshold voltages in nanometer scale ICs," *ICCD*, 2005, pp. 411-416.
[39] P. I. Pénzes and A. J. Martin, "Energy-delay efficiency of VLSI computations," *GLSVLSI*, 2002, pp. 104–111.
[40] H. P. Hofstee, "Power-constrained microprocessor design," *ICCD*, 2002, pp. 14–16.
[41] S. Borkar et al., "Parameter variations and impact on circuits and microarchitecture," *DAC*, 2003, pp. 338-342.
[42] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, Vol. 19, pp. 23-29, 1999.
[43] S. Zhang et al., "A probabilistic framework to estimate full-chip subthreshold leakage power distribution considering within-die and die-to-die P-T-V variations," *ISLPED*, 2004, pp. 156-161.
[44] M.M. Pelella, J.G. Fossum, and S. Krishnan, "Control of off-state current in scaled PD/SOI CMOS digital circuits," *International SOI conference*, 1998, pp. 147-148.
[45] S-C. Lin et al., "Analysis and implications of IC cooling for deep nanometer scale CMOS technologies", *IEDM*, 2005 (to appear).
[46] K. Banerjee, et al, "3-D ICs: A novel chip design for improving deep submicron interconnect performance and systems-on-chip integration," *Proceedings of the IEEE*, Vol. 89, pp. 602-633, 2001.
[47] N. Srivastava and K. Banerjee, "Performance analysis of carbon nanotube interconnects for VLSI applications", *ICCAD,* 2005, pp. 383-390.
[48] N. Srivastava, R. V. Joshi and K. Banerjee, "Carbon nanotube interconnects: implications for performance, power dissipation and thermal management", *IEDM*, 2005 (to appear).