Cool Chips: Opportunities and Implications for Power and Thermal Management

Sheng-Chih Lin, *Student Member, IEEE*, and Kaustav Banerjee, *Senior Member, IEEE*

***Abstract*—Alongside innovative device, circuit, and microarchitecture level techniques to alleviate power and thermal problems in nanoscale CMOS-based integrated circuits (ICs), chip cooling could be an effective knob for power and thermal management. This paper analyzes IC cooling while focusing on the practical temperature range of operation. Comprehensive analyses of chip cooling for various nanometer scale bulk-CMOS and Silicon-On-Insulator (SOI) technologies are presented. Unlike all previous works, this analysis employs a holistic approach (combines device, circuit and system level considerations) and also takes various electrothermal couplings between power dissipation, operating frequency and die temperature into account. While chip cooling always gives performance gain at the device and circuit level, it is shown that system level power defines a temperature limit beyond which cooling gives diminishing returns and an associated cost that may be prohibitive. A scaling analysis of this temperature limit is also presented. Furthermore, it is shown that on-chip thermal gradients cannot be mitigated by global chip cooling and that localized cooling can be more effective in removing hot-spots.**

***Index Terms*—Cooling, integrated circuits, performance, power consumption, thermal management.**

I. Introduction

FOR THE PAST 40 years, tracking Moore's Law [1] has been the goal of the semiconductor industry in the development of silicon integrated circuits. Shrinking transistor size with innovative technology provides significant benefits in the form of higher integration density, higher performance, and lower cost [2]. However, continuous scaling raises severe design challenges and concerns due to excessive power consumption (power density) and associated thermal problems, especially for high-performance microprocessors [3]–[5].

Table I summarizes key parameters predicted by the International Technology Roadmap for Semiconductors (ITRS) [6] for silicon technology in the near future. While switching energy per device decreases with scaling [Fig. 1(a)], preliminary calculation clearly shows that even if only switching power is taken into account, average power density continues to increase as illustrated in Fig. 1(b). Note that values (Trend 1) in Fig. 1(b) are derived based on Table I under the assumptions of: 1) maximum integration density, and 2) highest performance, which are not practical. However, the projected allowable maximum

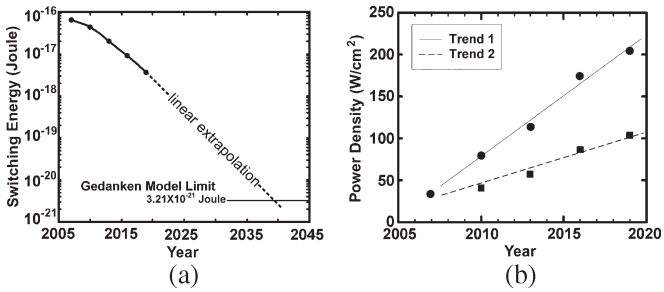


Fig. 1. (a) Trend of minimum transistor switching energy based on Table I. The fundamental lower limit of switching energy for irreversible logic computing is calculated using the gedanken model with the channel tunneling effect (see Appendix for more detail) [7]. (b) Trend of IC power density (Trend 1) with ITRS projected integration density and performance. Although switching energy per device decreases with scaling, switching power and density increases due to the fact that performance and packing density are also improved. It is assumed that 0.1% transistors switch simultaneously. Values shown here are exceptionally high (in reality, packing density and performance will not be as high as projected). Power density can be reduced (Trend 2) if the chip size is increased (to 620 mm^2 from the year 2010 and beyond) or if the switching activity is halved. However, this only mitigates the increase in power density. When the maximum allowable average power density is constrained by the limitation of heat removal capability as per ITRS projection, doubling transistor count with scaling requires innovative power and thermal management strategies.

TABLE I
HIGH-PERFORMANCE LOGIC TECHNOLOGY TREND TARGETS (ITRS 2006 EDITION) [6]

Year of Production	2007	2010	2013	2016	2019
	Planar Bulk		Double Gate		
Supply Voltage (V)	1.1	1.0	0.9	0.8	0.7
[1]Transistor (M)	1106	2212	4424	8848	17696
[2]Size (mm^2)	310	310	310	310	310
[3]L_g (nm)	25	18	13	9	6
[4]$I_{d,sat}$ (μA/μm)	1200	2050	2220	2713	2744
[5]$I_{sd,\ leakage}$ (μA/μm)	0.2	0.28	0.11	0.11	0.11
[6]Intrinsic Delay, τ (ps)	0.64	0.40	0.25	0.15	0.10
[7]Switching Energy (fJ)	0.0634	0.0447	0.0198	0.0091	0.0036
[8]Max. P.D. (W/cm^2)	60.97	63.87	63.87	63.87	63.87

[1]Functions per chip at production (million transistors)
[2]Chip size at production (mm^2)
[3]Physical gate length (nm)
[4]Effective saturation drive current (μA/μm)
[5]Source/Drain subthreshold off-state leakage current (μA/μm)
[6]Intrinsic transistor delay for NMOS devices at 25°C (ps)
[7]Energy per device switching transition with dimensions W/L_g=3 (fJ/device)
[8]Maximum allowable average power density (W/cm^2)

Manuscript received June 6, 2007. This work was supported in part by Intel Corporation, the University of California-MICRO program (03-004) and by the National Science Foundation (Award CCF-0541465). The review of this paper was arranged by Editor J. Welser.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: sclin@ece.ucsb.edu; kaustav@ece.ucsb.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2007.911763

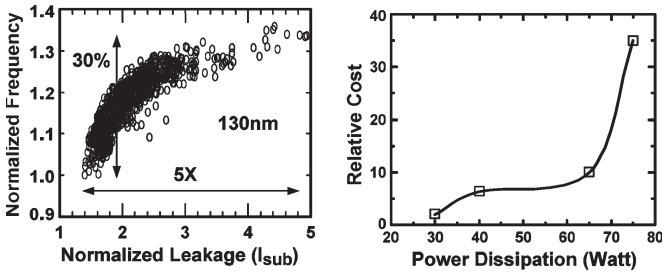


Fig. 2. (a) Distributions of frequency and standby leakage current for different microprocessors on a single wafer. (Courtesy: Intel) (b) Relative cost of heat removal from a microprocessor [11].

power density (shown in Table I) is approaching saturation in the next few years due to limited capability for system level heat removal.

In addition to switching power, thermal problems are exacerbated by the fact that leakage power forms a significant fraction of total chip power consumption [8]. Compared to switching power consumption, leakage power is undesirable and becoming a dominant factor limiting CMOS transistor scaling [4], [9].

Subthreshold leakage is the main leakage contributor in nanoscale CMOS and is highly temperature sensitive [10]. It increases rapidly with scaling due to continuous reduction in the supply voltage $(V_{\rm dd})$, which necessitates reduction of the threshold voltage $(V_{\rm th})$ to maintain required performance. Since power consumption is converted into heat, operating temperature rises, which significantly increases subthreshold leakage. Moreover, threshold voltage decreases with temperature and results in even higher subthreshold leakage. Furthermore, since the gap between the wavelength of light for optical lithography and the polysilicon gate length is increasing [3], transistor channel length exhibits a significant amount of variations, which further increases leakage power as shown in Fig. 2(a) [8].

A. Impact of Leakage on Chip Thermal Management

Increasing leakage power has significant implications for thermal management strategies (including packaging and cooling solutions) for nanometer scale ICs. Traditionally, the impact of power on package level thermal management strategies can be understood by

$$\theta_{\rm ja} = \frac{T_{\rm j} - T_{\rm amb}}{P_{\rm chip}} \tag{1}$$

where $\theta_{\rm ja}$ is a lumped value of the thermal resistance between the silicon junction to the ambient (environment outside the chip case). $P_{\rm chip}$ is total chip power consumption. $T_{\rm j}$ and $T_{\rm amb}$ are the average junction and ambient temperatures, respectively. Typically, a higher value of $\theta_{\rm ja}$ translates to a lower cost of the packaging and cooling solution. It can be observed that a larger value of $(T_{\rm j} - T_{\rm amb})$ allows a larger $\theta_{\rm ja}$ for dissipating the same power $(P_{\rm chip})$. In the recent past (for switching power dominant technologies), designers have allowed $T_{\rm j}$ to increase with increasing $P_{\rm chip}$, since maintaining larger $T_{\rm j}$ relaxes $\theta_{\rm ja}$ requirements. However, for deeply scaled technologies which are leakage dominant, larger $T_{\rm j}$ exponentially increases subthreshold leakage, thereby influencing $\theta_{\rm ja}$ and drastically increasing the cost of operation. The relationship between power dissipation and relative operating cost (including thermal management) is illustrated in Fig. 2(b).

B. Power Management via Device and Circuit Techniques

While continued scaling of CMOS technologies provides remarkable benefits in the form of higher transistor packing density, higher circuit performance, and lower cost of ICs, many leakage mechanisms of a nanoscale transistor become prominent [10], [12].

To suppress leakage, a significant amount of device level technology innovation and optimization has been applied. Short channel effects (SCE) [10], which lead to higher subthreshold leakage, have been shown to improve via substrate engineering. For instance, vertically nonuniform doping (retrograde channel profile) enhances inversion layer mobility because of lower surface doping [13], [14], while laterally nonuniform channel implants (halo doping) reduces threshold voltage roll-off by compensating 2-D charge-sharing effects in short channel transistors [15]–[17].

Another critical technology challenge is transistor gate tunneling leakage, which increases with ever-thinning silicon dioxide gate dielectric [18]. To alleviate this problem, a thicker insulating material with higher dielectric constant (High-κ) has been proposed to replace the thin silicon dioxide. Also, a metal gate electrode has been used to replace the polysilicon gate for better control of the threshold voltage [19], [20]. However, more investigations of this new structure are required, including tuning of appropriate metal gate work function, ensuring adequate channel mobility with high-κ, and ensuring gate dielectric reliability as well as good interface properties between insulator and semiconductor. Moreover, novel device structures (enhanced or beyond classical CMOS) including strained-Si, thin-body silicon-on-insulator , multigate devices (e.g., double-gate, FinFET, trigate, etc.), with better control of SCE have all been evaluated as attractive candidates as conventional bulk-CMOS approaches the scaling limit [6], [21]–[23].

Besides device level technologies, various circuit level power saving or management techniques are employed in state-of-the-art high-performance ICs. From a circuit point of view, switching and static power consumption can be modeled as

$$P_{\rm switching} = aC_{\rm eff}V_{\rm dd}^2F \tag{2}$$

$$P_{\rm leakage} = I_0{\rm e}^{-V_{\rm th}/\gamma S}(1-{\rm e}^{-V_{\rm ds}/\gamma S})W_{\rm eff}V_{\rm dd} \tag{3}$$

where a is the switching activity factor, $C_{\rm eff}$ accounts for total effective output-load capacitance of the circuit, and F denotes operating frequency. I_0 is the nominal leakage current, $W_{\rm eff}$ is effective transistor width (transistor width that contributes to the leakage current) of the entire chip, $V_{\rm ds}$ is drain to source voltage, and γ is a device parameter [10]. Subthreshold swing (S) is defined by (4) as the inverse of the slope of drain current $(I_{\rm ds})$ versus the gate to source voltage $(V_{\rm gs})$ characteristic curve

in subthreshold regime presented in a semilogarithmic plot. In (4), k is the Boltzmann constant, T is the temperature, and q is the electron charge

$$S = \left[\frac{\partial \log_{10}(I_{\rm ds})}{\partial V_{\rm gs}}\right]^{-1} \propto \frac{kT}{q}. \quad (4)$$

From (2) and (3), obviously, lowering $V_{\rm dd}$ and increasing $V_{\rm th}$ (through transistor channel length biasing and/or body biasing) can reduce power consumption. However, this requires trading-off performance [24] in conjunction with careful transistor sizing [25]. Relevant low-power design methodologies include dual—or multi—$V_{\rm dd}$ and $V_{\rm th}$ schemes as well as adaptive body biasing techniques [26]. Transistor gating is also considered for low-power or power-constrained designs. For instance, the clock gating technique is used to reduce clock tree power dissipation [27]. Power gating and sleep transistor insertion techniques reduce leakage by turning off idle circuitry [28]. At the architecture level, pipelining and parallel (including multicore) structures are often implemented in low-power designs. The throughput can be maintained at a lower $V_{\rm dd}$ by parallel implementation. Also, applying pipelining can reduce power consumption while the switching rate and $V_{\rm dd}$ are reduced [29]. However, these methods reduce power consumption at the cost of area, performance, or noise margin penalty.

As power and thermal problems become more critical with technology scaling, device and circuit level power reduction techniques alone may not suffice, particularly in light of the tradeoffs between various design metrics that such techniques necessitate. Hence, it is worthwhile to simultaneously explore other potential knobs beyond the existing or evolving array of device and circuit techniques for power and thermal management. As in many practical instances, we need an exponential to fight another exponential. Therefore, chip cooling could be an attractive option for leakage control and power/thermal management of high-performance ICs (as evident from (3) and (4) above).

C. Scope of this Work

Low temperature CMOS operation has been consistently studied for decades as a promising approach for improving performance due to higher carrier mobilities and for extending CMOS technology with steeper subthreshold slopes (smaller S) [30]–[36], especially in the range of sub-ambient temperatures. Prior work in chip cooling has primarily focused on sub-ambient [37], [38] or cryogenic [30], [31] temperatures (Fig. 3). However, the practicality of such operating temperatures is questionable. Moreover, these analyses were carried out at the device or circuit level only, without any system level considerations (including cooling power consumption). Although cooled chip operation can be expected to improve device and circuit level performance and reliability, there is no well-defined methodology which quantifies the real benefits of cooling in a holistic manner.

This paper presents a comprehensive evaluation of IC cooling applied to deeply scaled technologies in the range of practical temperatures (practical commercial range in Fig. 3) for high-performance ICs, including microprocessors [39]. In addition to device and circuit level implications, the benefits of cooling are also quantified from a system level power dissipation point of view. Moreover, various cooling options are discussed. Also, global and localized cooling are analyzed for hot-spot management.

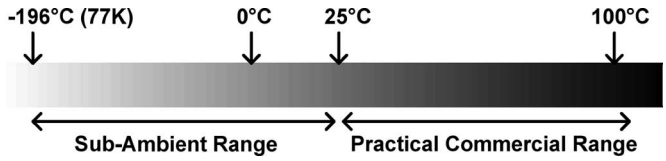


Fig. 3. Range of IC cooling. Typically, refrigeration or cryogenic cooling systems are required to achieve the temperatures below the ambient temperature (in the sub-ambient range). On the other hand, applying standard conduction, convection, or forced air cooling can only achieve the temperatures above the ambient temperature (in practical commercial range). In this paper, we mainly focus on the practical commercial range for chip cooling analysis.

The rest of this paper is organized as follows. Benefits of cooling at the device level are presented via detailed scaling analysis in Section II. This section includes: 1) comparisons between bulk and partially-depleted SOI (PD-SOI) type transistors in both ON and OFF states. 2) ratio of drive to leakage current $(I_{\rm on}/I_{\rm off})$ as a function of temperature at different technology nodes. 3) strategies for exploiting the benefit of cooling at the device level. In Section III, the impact of cooling at the circuit level is discussed. Analysis at the system level, including associated system power to quantify the real benefits of cooling, is illustrated in Section IV. In Section V, various cooling options are briefly discussed. A leakage-aware self-consistent substrate thermal profile estimation technique is then employed to compare the substrate temperature profile of a high-performance IC under global and localized cooling, and implications for hot-spot management are discussed. Finally, concluding remarks are made in Section VI.

II. Impact of Cooling at the Device Level

The primary motivation for employing cooling has been the increase in performance due to the improvement of carrier mobility. Mobility increases as temperature decreases mainly because of the reduction of carrier scattering caused by thermal vibrations of the semiconductor crystal lattice.

Transistor carrier mobility is a function of electric field, doping concentration, and temperature [10]. First-order carrier mobility (μ) in the saturation mode around room temperature can be modeled by (5), where n is around 1.3 and 1.2 for electrons and holes, respectively [32]. The mobility improvement is beneficial for enhancing CMOS performance.

$$\mu \propto T^{-n}. \quad (5)$$

Note that at cryogenic temperatures ($\leq$ 77 K), carrier mobility will be limited by impurity scattering. Thus, doping concentration needs to be considered in the mobility model for better accuracy [10]

As already shown in (4), the subthreshold swing (S) has the unit of millivolts per decade and represents sharpness

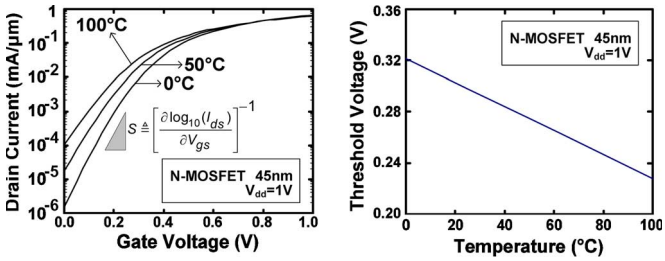


Fig. 4. (a) N-MOSFET $I_{\rm ds}-V_{\rm gs}$ curve (45 nm effective channel length) at different temperatures. (b) Device threshold voltage at different operating temperatures.

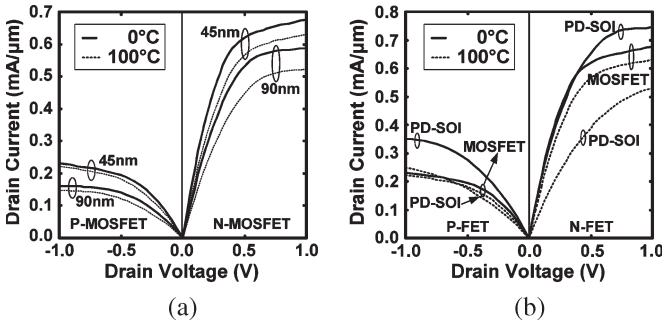


Fig. 5. Device $I-V$ characteristics at different temperatures. At the same operating temperature, P-FET drive current is lower than N-FET because the carrier (hole) mobility of P-FET is smaller than that (electron) of N-FET. Transistor drive current increases at lower operating temperatures. (a) Bulk MOSFETs at 45 and 90 nm effective channel lengths. (b) Bulk MOSFET versus PD-SOI with effective channel lengths of 45 and 120 nm, respectively.

of the transistor drain current transition between ON and OFF states. Typically, it is preferred to have a steep slope (i.e., a smaller value of S). Fig. 4(a) shows the $I_{\rm ds}-V_{\rm gs}$ curves for N-MOSFET at different operating temperatures (simulation based on [40] and [41]). The advantage of low temperature operation can be easily observed from (4) and Fig. 4(a) as S is directly proportional to temperature.

The threshold voltage $(V_{\rm th})$ of N-MOSFET is modeled by (6) when $V_{\rm ds}$ is small. Note that when $V_{\rm ds}$ is large, $V_{\rm th}$ decreases due to the effect of drain-induced barrier lowering. $V_{\rm fb}$ denotes the flat-band voltage, while $\Psi_{\rm B}$ is the potential difference between the Fermi level and intrinsic level. $\varepsilon_{\rm si}$ is silicon permittivity, $N_{\rm a}$ is the acceptor impurity density, and $C_{\rm ox}$ is the oxide capacitance per unit area [10]

$$V_{\rm th} = V_{\rm fb} + 2\Psi_{\rm B} + \frac{\sqrt{4\varepsilon_{\rm si} q N_{\rm a} \Psi_{\rm B}}}{C_{\rm ox}}. \tag{6}$$

In (6), both $V_{\rm fb}$ and $\Psi_{\rm B}$ are temperature dependent. In general, the temperature dependence of $V_{\rm th}$ is around -1 mV/K [10]. Fig. 4(b) illustrates the value of $V_{\rm th}$ at different operating temperatures (simulation based on [40] and [41]). Note that although $V_{\rm th}$ increases at lower temperatures and partially offsets the performance improvement resulting from the higher carrier mobility, the transistor on-current still increases at lower temperatures (Fig. 5).

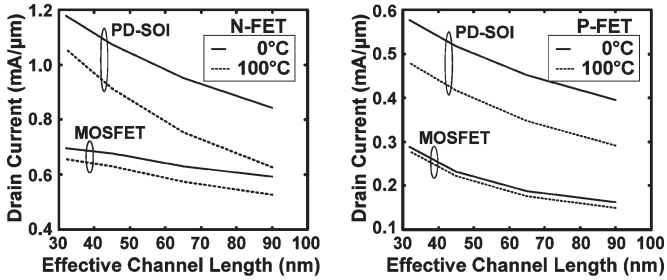


Fig. 6. Drive current for (a) N-FET and (b) P-FET as a function of effective channel length at 100 °C and 0 °C (cooled operation). Note that transistor drive current is the drain current at saturation mode. In general, drive current is expected to increase with technology scaling.

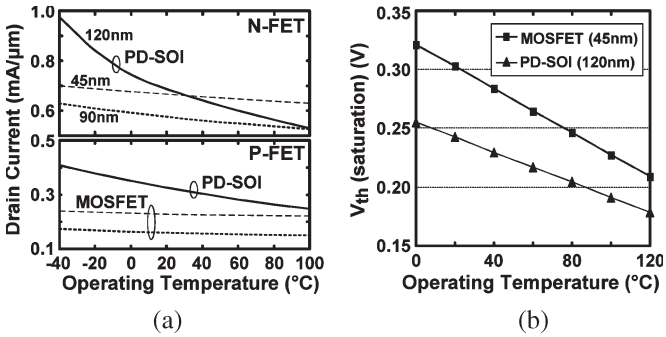


Fig. 7. (a) Drive (drain) current for N-FET (top) and P-FET (bottom) as a function of operating temperature, at different technology nodes. The floating-body PD-SOI type transistors show greater temperature sensitivity. (b) Rate of increase in saturated threshold voltage for bulk MOSFET (0.9 mV/°C) is larger than that of PD-SOI type transistor (0.6 mV/°C).

A. Enhancing Transistor Drive Current

The impact of cooling at the device level for bulk CMOS and floating-body PD-SOI type transistors is analyzed in the practical temperature range indicated in Fig. 3.

Thin-body SOI CMOS is an attractive alternative to bulk CMOS due to superior electrostatics and speed (lower junction capacitance), as well as lower subthreshold swing and latch-up/SER immunity. Fig. 5 compares the I–V characteristics of bulk and floating-body PD-SOI type transistors at different temperatures. Due to the smaller mobility of holes compared to that of electrons, the maximum drain current of P-type transistors are smaller than that of identical size N-type transistors.

Fig. 6 shows that drive (drain) current capability increases with transistor scaling. Due to higher carrier mobility at lower temperatures, it can be observed from Figs. 5 and 6 that higher drive current can be achieved by cooled operation across all technology nodes. In Fig. 7(a), while it is evident that drive current increases at lower temperatures for both bulk and PD-SOI, it is observed that SOI type transistors show greater sensitivity to temperature. This is due to the fact that the body to source voltage $(V_{\rm BS})$ of PD-SOI increases as temperature decreases [42], which causes a smaller increase in the saturated threshold voltage of PD-SOI type transistors compared to bulk transistors at lower temperatures, as shown in Fig. 7(b). Thus, the enhancement of drive current of PD-SOI transistors at lower temperatures is higher than that in bulk transistors.

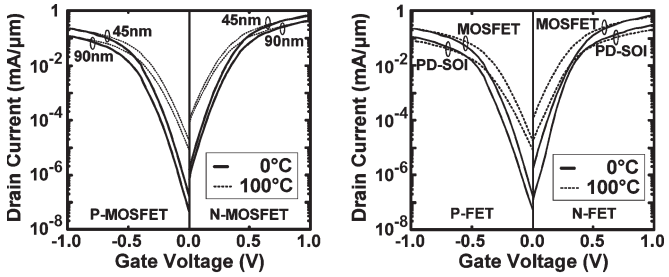


Fig. 8. Device subthreshold characteristics at different temperatures. (a) Bulk MOSFET for 45 nm ($V_{\rm dd} = 1.0$ V) and 90 nm ($V_{\rm dd} = 1.2$ V) effective channel length. (b) Bulk MOSFET versus PD-SOI with effective channel length of 45 nm ($V_{\rm dd} = 1.0$ V) and 120 nm ($V_{\rm dd} = 1.5$ V), respectively.

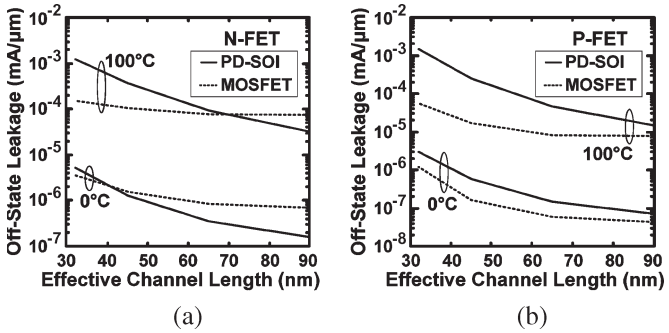


Fig. 9. Transistor OFF-state leakage current for (a) N-FET and (b) P-FET with shrinking device channel length. OFF-state leakage current decreases significantly under cooled operation.

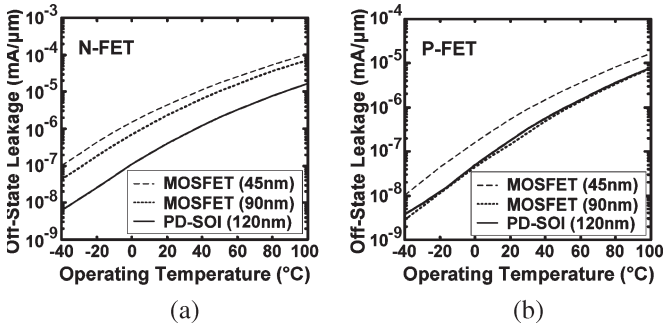


Fig. 10. OFF-state leakage current for (a) N-FET and (b) P-FET as a function of temperature at different technology nodes.

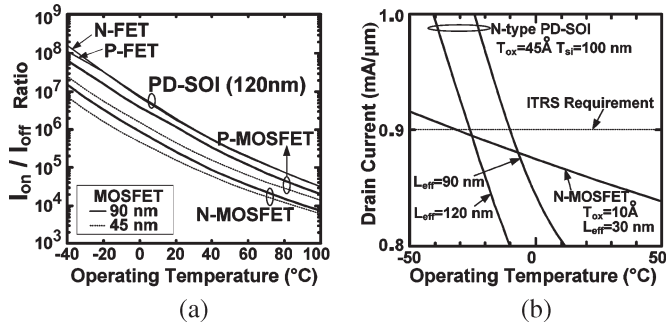


Fig. 11. (a) Ratio of drive current to leakage current as a function of temperature at different technology nodes. (b) ITRS requirement (nominal high-performance saturation drive current) for N-FET saturation drive current can be achieved by lowering the operating temperature to around −30 °C for N-MOSFET (30 nm effective channel length) without redesign. Curves of N-type PD-SOI devices (90 and 120 nm effective channel length) are also shown for comparison. As per Fig. 7 (a), lower amount of cooling is needed for the PD-SOI devices to meet the ITRS requirements.

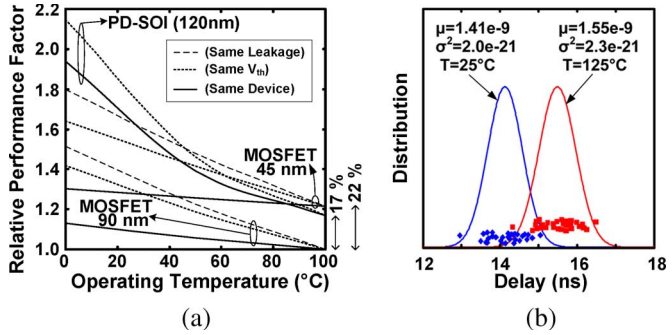


Fig. 12. (a) Relative performance factor (∝ transistor drive current) for three scenarios (same device, same $V_{\rm th}$, and same leakage) at different operating temperatures for PD-SOI (120 nm effective channel length), and N-MOSFET (45 nm and 90 nm effective channel length). Curves are normalized to the values of each device at 100 °C. The relative performance factor of N-MOSFET with 45 nm effective channel length and PD-SOI with 120 nm effective channel length at 100 °C are 22% and 17% higher than that of the N-MOSFET with 90 nm effective channel length. (b) Monte Carlo analysis of the propagation delay of a 9-stage inverter chain. The distribution of the delay is indicated by "♦" and "■" for 25 °C and 125 °C respectively.

B. Reducing Transistor Leakage

The device subthreshold characteristics for both bulk and SOI type transistors are shown in Fig. 8. The device under cooled operation clearly exhibits a steeper subthreshold slope than that under normal operation. It can be observed from Fig. 9 that subthreshold leakage increases exponentially with transistor scaling, while Fig. 10 shows that the leakage decreases exponentially with lower operating temperature. Hence, cooling can very effectively offset the increase in leakage current due to technology scaling, which is desirable for improving performance. Thus, as shown in Fig. 11(a), lowering of temperature significantly improves the $I_{\rm on}$ to $I_{\rm off}$ ratio for bulk as well as for SOI devices at all technology nodes. Note that the P-MOSFET has higher ratio than the N-MOSFET. This is due to the fact that at the same technology node, $I_{\rm off}$ is much lower for P-MOSFET as shown in Fig. 8(a), which is plotted using a logarithmic scale. Fig. 11(b) shows the amount of cooling that will be needed to meet the ITRS prescribed requirement for saturation drive current without redesigning.

C. Exploiting the Benefits of Cooling at the Device Level

Fig. 12(a) summarizes the relative improvement in performance as a result of cooling for different design strategies. The benefit of cooling alone can be seen in the scenario where no redesign is employed (same device). It is important to note that although the threshold voltage increases at lower operating temperatures and partially offsets performance improvement gained from higher carrier mobility, cooling still provides a net improvement in performance. Additionally, redesign strategies can be used to further enhance the benefit of cooling. For instance, adjusting threshold voltage by body-biasing (e.g., applying forward body bias to lower threshold voltage in N-MOSFET) to maintain the same threshold voltage (same $V_{\rm th}$) at lower temperatures shows a higher net performance improvement mainly due to higher mobility. Further lowering

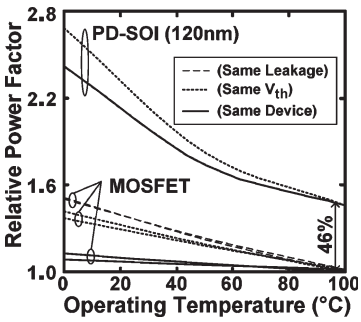


Fig. 13. Relative power factor (calculated based on switching power dissipation) for three scenarios at different operating temperatures, normalized to the values at 100 °C. Note that only two scenarios (same device and same $V_{\rm th}$) are shown for the PD-SOI case. Three scenarios are shown for N-MOSFET. The upper lines (higher relative power factor) correspond to N-MOSFET with 45 nm effective channel length and the lower lines are for 90 nm. The relative power factor of PD-SOI with 120 nm effective channel length at 100 °C is 46% higher than that of the N-MOSFETs (with 45 and 90 nm effective channel lengths).

of the threshold voltage to maintain the same OFF-state leakage current (same leakage) at lower temperatures, results in maximum improvement in performance that can be achieved by cooling. As shown in Fig. 12(a), redesign strategies become increasingly effective in exploiting the benefits from cooling as technology scales.

III. Impact of Cooling at the Circuit Level

Higher device drive (drain) current capability at lower temperature enhances circuit performance. The distribution of gate propagation delay of a nine-stage inverter chain (30 samples) under different operating temperatures was estimated by Monte Carlo analysis. As shown in Fig. 12(b), the mean value (μ) of gate propagation delay improved by 9% at the lower temperature. Moreover, variation in the circuit performance due to channel length variation can also be mitigated. (σ^2 reduces by 13% from 125 °C to 25 °C). Furthermore, cooled operation benefits back-end performance and reliability. Lower operating temperatures lead to smaller wire resistance per unit length, which reduces delay in signal lines and static IR-drop in power/ground networks [43]. Reliability of interconnects [electromigration (EM)] [44] and inter-layer dielectrics (Time-Dependent Dielectric Breakdown) [10] also improves due to cooling. For semiglobal and global wires, more aggressive interconnect scaling (narrower width with fixed or narrower spacing) can be allowed under cooled operation without degrading RC delay and EM reliability. This would also improve the wireability of the chip. On the other hand, scaled wires can offset any inductive effects that may become prominent due to reduced resistance per unit length at lower temperatures [45]. Thus interconnect scaling and cooled operation can be mutually beneficial. Also, at lower temperatures, intra-wire capacitance per unit length can be reduced significantly for smaller aspect ratio wires (reduced metal thickness with constant width) while maintaining the same resistance per unit length. This will lower the delay per unit length, and thereby enhance the rate at which bits can be transmitted per unit chip edge, i.e., bandwidth [46]. Lower delay per unit length will also reduce the number and size of repeaters needed along global interconnects, which will lead to lower power dissipation [47]. Benefits of cooling on front-end and back-end reliability have been recently demonstrated in [48].

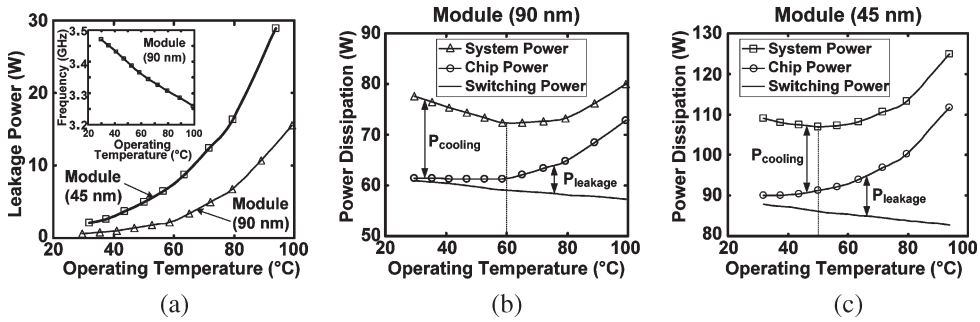
Fig. 13 shows the increase of active (switching) power dissipation (due to higher frequency of operation) at different operating temperatures for the same design scenarios analyzed in Fig. 12(a). Although lower temperature enhances the performance and reliability at the device and circuit level, the improvement in performance comes at the cost of increasing power dissipation. Moreover, it is inadequate to analyze the benefit of cooling while simply considering switching $(P_{\rm switching})$ and leakage $(P_{\rm leakage})$ power dissipation at the device level and ignoring the additional cooling power $(P_{\rm cooling})$ required to achieve lower junction temperature. The aspect of system level power dissipation is addressed in the following section.

IV. Analysis of Cooling at the System Level

A variety of cooling technologies have been proposed for improving high-performance ICs [49]–[58]. It has been shown that system performance cannot be correctly evaluated without considering electrothermal couplings, junction temperature, and associated cooling power [59]. The inset of Fig. 14(a) shows the improvement of electrothermally-aware system level performance while applying cooling. As expected, similar to analyses at the device and circuit level, system performance will always improve under cooled operation. However, considering improvements in performance alone cannot quantify the real benefits of cooling—associated system power must be taken into account. As a result, existing metrics in the literature that trade-off power and performance, such as energy-delay-product and MIPS/watt, may not be meaningful for leakage-dominant technologies if electrothermal couplings are neglected, which translates to neglecting the dynamic nature of system performance improvement with power.

Fig. 14(a) shows leakage power dissipation for two identical test microprocessor designs at different technology nodes under the application of active cooling, using the methodology described in [59]. It can be observed that leakage power dissipation decreases significantly as more cooling power is applied and the reduction of leakage power (leakage power reduction per °C) becomes greater as technology scales.

Fig. 14(b) and (c) show that chip power $(P_{\rm chip} = P_{\rm switching} + P_{\rm leakage})$ decreases as more cooling is applied, mainly as a result of decreasing leakage power as shown in Fig. 14(a). The total system power $(P_{\rm system} = P_{\rm chip} + P_{\rm cooling})$, however, decreases only as long as the savings in chip power dissipation remains greater than the additional power required for cooling. Note that $P_{\rm cooling} = (1-\eta)\cdot P_{\rm chip}$, where η is the cooling efficiency $(\eta < 1)$. Hence, it can be observed from Fig. 14(b) and (c) that there is a clear minimum point in the curve of total system power $(P_{\rm system})$ that determines the practical limit of operating temperature $(T_{\rm limit})$, beyond which further cooling does not lead to any overall system power savings. This limit occurs at an increasingly lower temperature as technology scales (from 90 to 45 nm). Since performance increases at lower operating temperatures [shown in the inset



(a) (b) (c)

Fig. 14. Electrothermally-aware system level evaluation of power dissipation. A minimum $P_{\rm system}$ determines the practical limit beyond which further cooling does not lead to any power saving. (a) $P_{\rm leakage}$ as a function of operating temperature. The inset shows that chip frequency increases as operating temperature decreases. (b) Minimum $P_{\rm system}$ (90 nm) is around 60 °C ($T_{\rm limit} = 60$ °C). (c) Minimum $P_{\rm system}$ (45 nm) is around 50 °C ($T_{\rm limit} = 50$ °C).

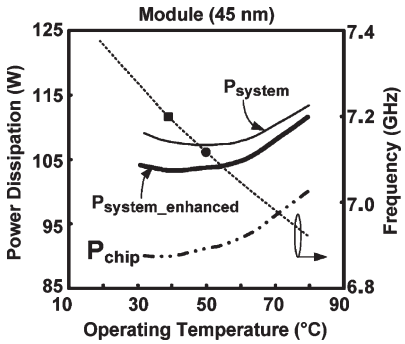


Fig. 15. Electrothermally-aware system level evaluation of power dissipation for 45 nm module. $P_{\rm system}$ and $P_{\rm chip}$ curves are from Fig. 14 (c). The system power with 70% enhanced cooling efficiency ($P_{\rm system\_enhanced}$) is also shown. In addition, the frequency versus operating temperature curve for 45 nm module is superimposed and plotted using the second y-axis. "●" denotes the $T_{\rm limit}$ from Fig. 14 (c) and "■" denotes the $T_{\rm limit}$ with enhanced cooling efficiency. The minimum point of $P_{\rm system\_enhanced}$ moves toward a lower temperature which leads to higher performance.

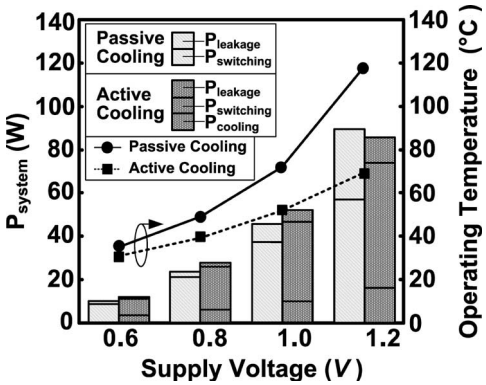


Fig. 16. Electrothermally-aware system level evaluation of power dissipation and operating temperature of 100 nm Module. Four different supply voltage scenarios (0.6, 0.8, 1.0, 1.2 V) are demonstrated under passive and active cooling. $P_{\rm system}$ for the case of passive cooling (left bar) has a detailed breakdown of the power consumption ($P_{\rm leakage}$ and $P_{\rm switching}$) while the case of active cooling (right bar) has a detailed breakdown of power ($P_{\rm leakage}$, $P_{\rm switching}$, and $P_{\rm cooling}$). For the case of $V_{\rm dd} = 1.2$ V, $P_{\rm system}$ for active cooling is less than that for passive cooling.

of Fig. 14(a)], a lower practical limit ($T_{\rm limit}$) indicates that as technology scales, the benefit that can be derived from cooling increases.

Fig. 15 demonstrates that the practical limit of cooling ($T_{\rm limit}$) can be further extended toward a lower operating temperature (and hence higher performance) by enhancing cooling efficiency.

Design parameters such as supply voltage can also affect the usefulness of cooling in improving chip performance and power dissipation. Fig. 16 compares the total system power dissipation ($P_{\rm system}$) for the case where no cooling power is applied (passive cooling) to the case where additional cooling power is used (active cooling) to lower the operating temperature. Detailed power breakdown for each case is also shown for comparison. As expected, at lower supply voltages, although active cooling results in lower operating temperature, the total system power dissipation is higher than that with passive cooling. However, at higher supply voltage (e.g., $V_{\rm dd} = 1.2$ V) active cooling not only leads to lower operating temperature, but also results in lower total system power consumption compared to that under passive cooling. This is because the extra power spent on cooling is lower than the corresponding saving in leakage power.

TABLE II
SUMMARY OF MAJOR ACTIVE COOLING OPTIONS [49]

Technology	Advantages	Disadvantages
Fan sinks + heat pipe (hybrid)	Compact, versatile	Reliability, space, limited to ambient temperature
Thermoelectric	Spot cooling	Reliability, low capacity
Liquid cooling	High surface heat transfer	Sealing, cost, maintenance, packaging
Direct immersion	High capacity	Cost, sealing, packaging
Refrigeration	Sub-ambient	Cost, power, space, packaging
Cryogenics	Super cooling	Cost, power, packaging

V. OPPORTUNITIES FOR THERMAL MANAGEMENT

A. Cooling Techniques

Cooling techniques can be broadly classified into two types based on cooling power consumption. Passive cooling ($P_{\rm cooling} = 0$) denotes cooling by conduction (heat sink) and/or natural convection by air, while active cooling represents different types of cooling schemes with associated external cooling power ($P_{\rm cooling} > 0$). Typically, pure passive cooling is only applicable for systems with low power consumption due to its low heat removal capability, which is limited by the heat sink (size, thermal conductivity, etc.) and surrounding temperature. Depending on the application, different active cooling technologies have been explored and evaluated. Table II

summarizes the pros and cons for various major active cooling options [49].

Cooling techniques for sub-ambient temperature operation, including refrigeration and cryogenics, are only applicable for specialized use to achieve required performance when cost and cooling power consumption are not the primary concerns. The efficiency of various refrigeration techniques can be compared by the coefficient of performance (COP), which is defined as the ratio of cooling capacity to power consumption by the refrigerator ($\mathrm{COP} = Q_{\mathrm{cooling}}/W_{\mathrm{power}}$, where Q_{cooling} is the cooling capacity and W_{power} is the power consumption by the refrigerator). In [60], various refrigeration techniques are listed and discussed.

For operations above the ambient temperature, among the major cooling options listed in Table II, hybrid cooling (involving fan, heatsink, and heat pipe) is by far the most commonly employed approach for current commercial high-performance ICs, including microprocessors, because of the cost-driven market. The heat removal capability of air cooling is mainly determined by Newton's law of cooling as shown in (7), where Q denotes the heat flow (heat removal in watt), h is the convection heat transfer coefficient, A is the surface area, T_{surface} denotes the surface temperature, and T_{f} is the fluid (air) temperature

$$Q = hA(T_{\mathrm{surface}} - T_{\mathrm{f}}). \tag{7}$$

Heat removal capability by air (hybrid) cooling can be enhanced by increasing h, A and T_{surface} with efficient forced convection (fan), novel heatsink design, and better packaging materials with low thermal resistance, respectively.

Moreover, the limit of heat removal capability (Q) can be further improved by using liquid as the coolant (water-cooled heatsink). The concept of microchannel cooling with liquid has been introduced and investigated for applications with higher heat removal requirements (e.g., large size or array of high performance ICs). In [51], experimentally, a thermal resistance of 0.09 °C/W is demonstrated over a uniform area of 1 cm^2 with power density of 790 W/cm^2 (the substrate temperature rise was 71 °C above the input liquid temperature of 23 °C). Microchannels with single- and two-phase cooling are also investigated and discussed in [52]. While two-phase cooling can potentially achieve lower thermal resistance and better heat removal capacity, this technology is still under development.

However, due to the presence of hot-spots (larger thermal gradient) [3] on the substrate of high-performance ICs, the need for localized (selective) cooling necessitates innovative microcoolers such as solid-state thin-film thermoelectric coolers (TEC) and ionic wind engines [61]. Recently, thin-film TEC is becoming an attractive option mainly due to its compact structure and larger cooling capability [55]–[58].

B. Effective Cooling Strategy for Hot-Spot Management

To comprehend the impact of cooling on thermal gradients and hot-spots, a self-consistent electrothermal substrate thermal profile generating methodology has been developed (Fig. 17) [62], [63]. Thermal parameters of a typical microprocessor package structure, Flip-Chip Land-Grid-Array and a socket that interfaces with the printed-circuit board were used for the thermal profile estimation. The package thermal model used in this analysis and thermal parameters of different packaging layers are listed in [63]. This methodology is based on solutions of the parabolic heat partial differential equations (both vertical and lateral heat transfer are considered) incorporating an electrothermally-aware self-consistent approach [59]. The parabolic PDEs were solved using the Alternating-Direction-Implicit method [64] for achieving high computation efficiency.

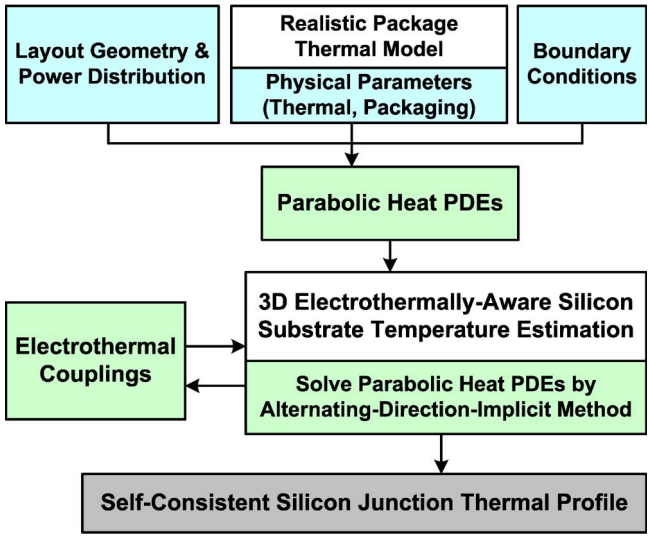


Fig. 17. Overview of the electrothermally-aware substrate thermal profile generator.

An example chip design (die size: 10 mm × 10 mm) with power densities per functional block is shown in Fig. 18(a). The substrate temperature profile, Fig. 18(b), shows several hot-spots and the highest junction temperature is around 73 °C. Although the results shown here are specific to the aforementioned IC, the conclusions drawn are more generic. Fig. 19 shows the effect of applying global and localized cooling strategies on hot-spot management. As shown in Fig. 19(a), a lower junction-to-ambient thermal resistance (θ_{ja}) obtained by applying global cooling (through better interface material, higher cooling efficiency, etc.) reduces the maximum junction temperature. However, on-chip hot-spots and thermal gradients still remain. On the other hand, localized cooling solutions such as local spray cooling and thin-film thermoelectric coolers can be applied to eliminate the hot-spots. For example, if a thin-film thermoelectric cooler is placed on the backside of the wafer below the location of the bottom-right hot-spot, it can effectively eliminate the targeted hot-spot as shown in Fig. 19(b).

VI. Conclusion

In conclusion, a comprehensive analysis of chip cooling for various nanometer scale bulk-CMOS and SOI technologies combining device, circuit and system level considerations along with electrothermal couplings between power, frequency and die temperature has been presented. At the device level, it is shown that floating-body PD-SOI based technologies are more responsive to cooling. It is also demonstrated that lowering the operating temperature in leakage-dominant nanometer scale CMOS technologies can reduce overall cost, since the power

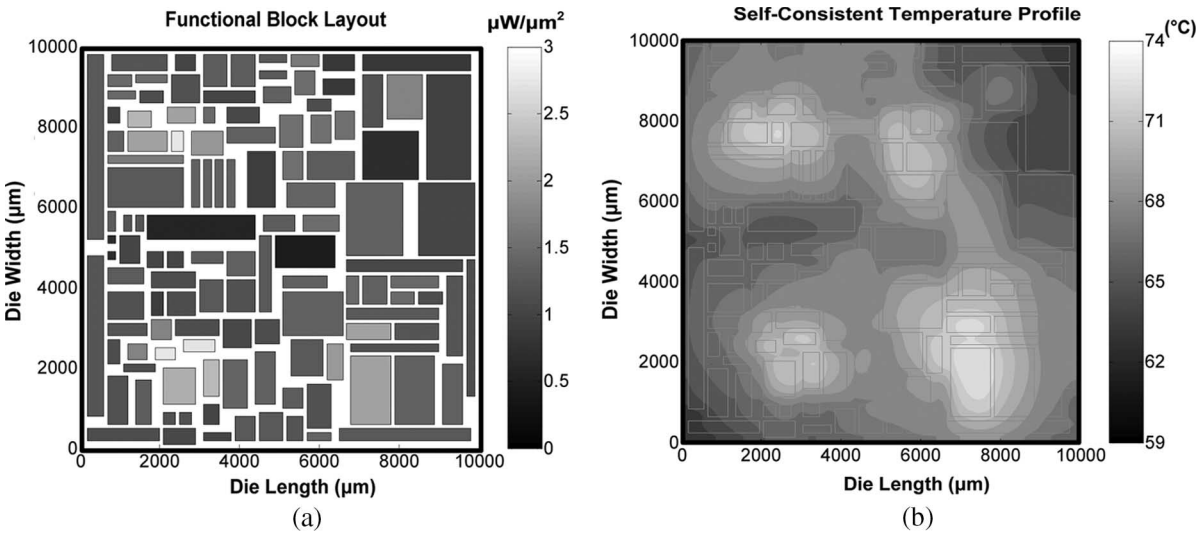


Fig. 18. (a) Functional block layout of a test chip showing power density associated with each block. Nominal total power consumption is 90 W. (b) Spatial substrate temperature profile of the test chip generated using the methodology described in Fig. 17. Four hot-spots can be observed. The highest temperature ($T_{\max}$) is around 73 °C.

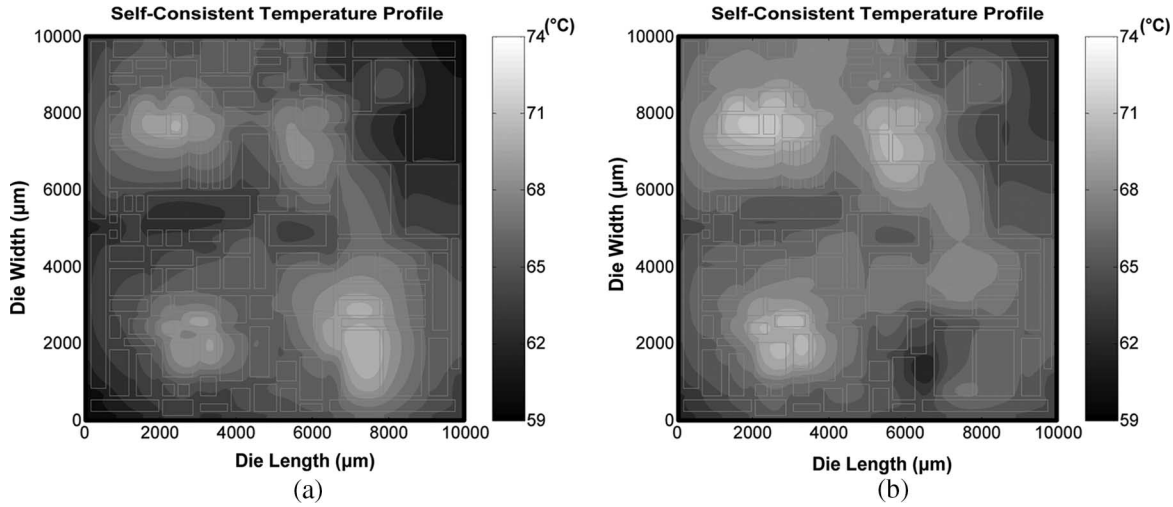


Fig. 19. (a) Spatial substrate temperature profile of the test chip generated using the methodology described in Fig. 17. It is assumed that the thermal resistance of all packaging layers reduces by 20% under global cooling. Although $T_{\max}$ (70 °C) decreases, the hot-spots remain. (b) Temperature profile of the test chip after integrating a thin-film thermoelectric cooler at the bottom-right hot-spot. Now, only three hot-spots can be observed.

needed for cooling may be regained from the lower leakage of the cooled devices. However, while cooling always gives performance gains at the device and circuit level, considering system level power consumption can clearly identify a temperature limit beyond which cooling gives diminishing returns. Also, the benefit that can be derived from cooling increases as technology scales. Finally, it is shown that localized cooling will be more effective for hot-spot management.

APPENDIX

The expression for the lower limit of switching energy considering both classic and quantum transport phenomena is summarized below (based on [7]) for the convenience of the readers.

Having "distinguishable states" and capability for a "conditional change of state" with a physical carrier are the two fundamental properties of a material subsystem for representing classical binary information. For instance, the field-effect transistor can be the subsystem where two wells (source and drain) are separated by an energy barrier (channel). As shown in Fig. 20, the electron can change state via two mechanisms: classic over-barrier and quantum mechanical tunneling transitions.

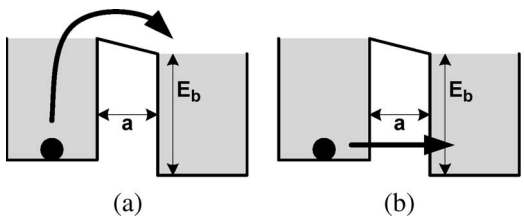


Fig. 20. Mechanisms for an electron to change states. $E_{\rm b}$ is the energy barrier and a is the barrier width (transistor channel length). (a) Classic over-barrier transition. (b) Quantum mechanical tunneling transition.

When only over-barrier transition is considered, the well-known Shannon–von Neumann–Landauer (SNL) energy limit

per switching (E_{SNL}) is the solution of the following equation where Π_{classic} denotes the probability of transitions between two wells, E_{b} is the barrier energy, k is the Boltzmann constant, and T is the temperature. Note that the distinguishability is lost when the probability (Π_{classic}) is 50% (i.e., $\phi_{\mathrm{classic}} = 0.5$)

$$\prod_{\mathrm{classic}} = \exp\left(-\frac{E_{\mathrm{b}}}{kT}\right) = 0.5 \Rightarrow E_{\mathrm{b}} = E_{\mathrm{SNL}} = kT(\ln 2). \quad \text{(A1)}$$

However, when barrier width (a) is small, the electron can pass through the barrier by tunneling even if the energy of the particle is less than E_{b}. The probability of tunneling (Π_{quantum}) is given by the Wentzel–Kramers–Brillouin approximation [65]

$$\prod_{\mathrm{quantum}} = \exp\left(-\frac{2\sqrt{2m}}{\hbar} a \sqrt{E_{\mathrm{b}}}\right) \quad \text{(A2)}$$

where m denotes the effective mass of the particle, and $\hbar$ is the Dirac constant (reduced Planck constant).

Therefore, the minimal energy per switch at the limits of distinguishability including classic and quantum transports can be solved when the total probability (Π_{error}) of transitions equals 50% (0.5) as shown in

$$\prod_{\mathrm{error}} = \prod_{\mathrm{classic}} + \prod_{\mathrm{quatum}} - \left(\prod_{\mathrm{classic}} \bullet \prod_{\mathrm{quantum}}\right) = 0.5. \quad \text{(A3)}$$

An approximate solution for (A3) is shown in (A4) and the minimal energy ($E_{\mathrm{b}}^{\mathrm{min}}$) is around $3.21 \cdot 10^{-21}$ J at 300 K

$$E_{\mathrm{b}}^{\mathrm{min}} \cong kT(\ln 2) + \frac{\hbar^2 (\ln 2)^2}{8ma^2}. \quad \text{(A4)}$$

ACKNOWLEDGMENT

The authors would like to thank Dr. G. Chrysler, Dr. R. Mahajan, and Dr. V. De from Intel Corporation for useful discussions and suggestions.

REFERENCES

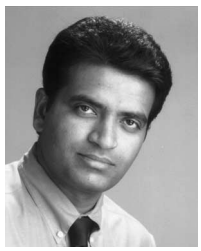
[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.
[2] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SSC-9, no. 5, pp. 256–268, Oct. 1974.
[3] P. Gelsinger, "Gigascale Integration for Teraops Performance–Challenges, Opportunities, and New Frontiers," in *Proc. 41st DAC Keynote*.
[4] D. J. Frank, "Power-constrained CMOS scaling limits," *IBM J. Res. Develop.*, vol. 46, no. 2/3, pp. 235–244, 2002.
[5] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, no. 4, pp. 23–29, Jul./Aug. 1999.
[6] *International Technology Roadmap for Semiconductors (ITRS)*. [Online]. Available: http://www.itrs.net
[7] V. V. Zhirnov, R. K. Cavin, III, J. A. Hutchby, and G. I. Bourianoff, "Limits to binary logic switch scaling—A Gedanken model," *Proc. IEEE*, vol. 91, no. 11, pp. 1934–1939, Nov. 2003.
[8] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proc. DAC*, 2003, pp. 338–342.
[9] Y. Taur, "CMOS design near the limit of scaling," *IBM J. Res. Develop.*, vol. 46, no. 2/3, pp. 213–222, 2002.
[10] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[11] *International Electronics Manufacturing Initiative (iNEMI) 2004 Thermal Roadmap*.
[12] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.
[13] H. S. P. Wong, D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. J. Welser, "Nanoscale CMOS," *Proc. IEEE*, vol. 87, no. 4, pp. 537–570, Apr. 1999.
[14] I. De and C. M. Osburn, "Impact of super-steep-retrograde channel doping profiles on the performance of scaled devices," *IEEE Trans. Electron Devices*, vol. 46, no. 8, pp. 1711–1717, Aug. 1999.
[15] C. F. Codella and S. Ogura, "Halo doping effects in submicron DI-LDD device design," in *IEDM Tech. Dig.*, 1985, pp. 230–233.
[16] G. G. Shahidi, J. Warnock, S. Fischer, P. A. McFarland, A. Acovic, S. Subbanna, E. Ganin, E. Crabbe, J. Comfort, J. Y.-C. Sun, T. H. Ning, and B. Davari, "High-performance devices for a 0.15-μm CMOS technology," *IEEE Electron Device Lett.*, vol. 14, no. 10, pp. 466–468, Oct. 1993.
[17] L. Su, S. Subbanna, E. Crabbe, P. Agnello, E. Nowak, R. Schulz, S. Rauch, H. Ng, T. Newman, A. Ray, M. Hargrove, A. Acovic, J. Snare, S. Crowder, B. Chen, J. Sun, and B. Davari, "A high-performance 0.08-μm CMOS," in *VLSI Symp. Tech. Dig.*, 1996, pp. 12–13.
[18] S. H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFETs," *IEEE Electron Device Lett.*, vol. 18, no. 5, pp. 209–211, May 1997.
[19] R. Chau, J. Brask, S. Datta, G. Dewey, M. Doczy, B. Doyle, J. Kavalieros, B. Jin, M. Metz, A. Majumdar, and M. Radosavljevic, "Application of high-κ gate dielectrics and metal gate electrodes to enable silicon and non-silicon logic nanotechnology," *Microelectron. Eng.*, vol. 80, pp. 1–6, Jun. 2005.
[20] E. P. Gusev, V. Narayanan, and M. M. Frank, "Advanced high-κ dielectric stacks with PolySi and metal gates: Recent progress and current challenges," *IBM J. Res. Develop.*, vol. 50, no. 4/5, pp. 387–410, 2006.
[21] G. G. Shahidi, "SOI technology for the GHz era," *IBM J. Res. Develop.*, vol. 46, no. 2/3, pp. 121–131, 2002.
[22] D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: How short can Si go?" in *IEDM Tech. Dig.,* 1992, pp. 553–556.
[23] D. Hisamoto, W. C. Lee, J. Kedzierski, E. Anderson, H. Takeuchi, K. Asano, T. J. King, J. Bokor, and C. Hu, "A folded-channel MOSFET for deep-sub-tenth micron era," in *IEDM Tech. Dig.*, 1998, pp. 1032–1034.
[24] S.-C. Lin, N. Srivastava, and K. Banerjee, "A thermally aware methodology for design-specific optimization of supply and threshold voltages in nanometer scale ICs," in *Proc. ICCD*, 2005, pp. 411–416.
[25] D. Markoviæ, V. Stojanoviæ, B. Nikoliæ, M. A. Horowitz, and R. W. Brodersen, "Methods for true energy-performance optimization," *IEEE J. Solid-State Circuits*, vol. 39, no. 8, pp. 1282–1293, Aug. 2004.
[26] J. W. Tschanz, S. Narendra, R. Nair, and V. De, "Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors," *IEEE J. Solid-State Circuits*, vol. 38, no. 5, pp. 826–829, May 2003.
[27] M. Pedram and J. Rabaey, *Power Aware Design Methodologies*. Norwell, MA: Kluwer, 2002.
[28] S. Mutah *et al.*, "1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS," *IEEE J. Solid-State Circuits*, vol. 30, no. 8, pp. 847–854, Aug. 1995.
[29] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 474–484, Apr. 1992.
[30] F. H. Gaensslen, V. L. Rideout, E. J. Walker, and J. J. Walker, "Very small MOSFETs for low temperature operation," *IEEE Trans. Electron Devices*, vol. ED-24, no. 3, pp. 218–229, Mar. 1977.
[31] J. Y. C. Sun, Y. Taur, R. H. Dennard, and S. P. Klepner, "Submicrometer-channel CMOS for low-temperature operation," *IEEE Trans. Electron Devices*, vol. ED-34, no. 1, pp. 19–27, Jan. 1987.
[32] W. F. Clark, B. El-Kareh, R. G. Pires, S. L. Titcomb, and R. L. Anderson, "Low temperature CMOS—A brief review," *IEEE Trans. Compon., Hybrids, Manuf. Technol.*, vol. 15, no. 3, pp. 397–404, Jun. 1992.
[33] J. D. Plummer, "Low temperature CMOS devices and technology," in *IEDM Tech. Dig.*, 1986, pp. 378–381.
[34] Y. Taur and E. J. Nowak, "CMOS devices below 0.1 μm: How high will performance go?" in *IEDM Tech. Dig.,* 1997, pp. 215–218.
[35] S. J. Wind, L. Shi, K. L. Lee, R. A. Roy, Y. Zhang, E. Sikorski, P. Kozlowski, C. D'emic, J. J. Bucchignano, H. J. Wann,

R. G. Viswanathan, J. Cai, and Y. Taur, "Very high performance 50 nm CMOS at low temperature," in *IEDM Tech. Dig.*, 1999, pp. 928–930.
[36] S. Takagi, M. Iwase, and A. Toriumi, "On the universality of inversion-layer mobility in N- and P-channel MOSFETs," in *IEDM Tech. Dig.*, 1988, pp. 398–401.
[37] B. Yu, H. Wang, C. Riccobene, H.-S. Kim, Q. Xiang, M.-R. Lin, L. Chang, and C. Hu, "Nanoscale CMOS at low temperature: Design, reliability, and scaling trend," in *Proc. Int. Symp. VLSI-TSA*, 2001, pp. 23–25.
[38] I. Aller, K. Bernstein, U. Ghoshal, H. Schettler, S. Schuster, Y. Taur, and O. Torreiter, "CMOS circuit technology for sub-ambient temperature operation," in *Proc. IEEE ISSCC*, 2000, pp. 214–215.
[39] S.-C. Lin, R. Mahajan, V. De, and K. Banerjee, "Analysis and implications of IC cooling for deep nanometer scale CMOS technologies," in *IEDM Tech. Dig.*, 2005, pp. 1041–1044.
[40] *Predictive Technology Model (PTM)*. [Online]. Available: http://www.eas.asu.edu/~ptm
[41] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit design," in *Proc. IEEE CICC*, 2000, pp. 201–204.
[42] M. M. Pelella, J. G. Fossum, and S. Krishnan, "Control of off-state current in scaled PD/SOI CMOS digital circuits," in *Proc. IEEE Int. SOI Conf.*, 1998, pp. 147–148.
[43] A. H. Ajami, K. Banerjee, and M. Pedram, "Scaling analysis of on-chip power grid voltage variations in nanometer scale ULSI," *Int. J. Analog Integr. Circuits Signal Process.*, vol. 42, no. 3, pp. 277–290, Mar. 2005.
[44] K. Banerjee and A. Mehrotra, "Global (interconnect) warming," *IEEE Circuits Devices Mag.*, vol. 17, no. 5, pp. 16–32, Sep. 2001.
[45] K. Banerjee and A. Mehrotra, "Analysis of on-chip inductance effects for distributed RLC interconnects," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 8, pp. 904–915, Aug. 2002.
[46] M. L. Mui, K. Banerjee, and A. Mehrotra, "A global interconnect optimization scheme for nanometer scale VLSI with implications for latency, bandwidth and power dissipation," *IEEE Trans. Electron Devices*, vol. 51, no. 2, pp. 195–203, Feb. 2004.
[47] K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Trans. Electron Devices*, vol. 49, no. 11, pp. 2001–2007, Nov. 2002.
[48] A. Hokazono, S. Kawanaka, K. Tsumura, Y. Hayashi, H. Tanimoto, T. Enda, N. Aoki, K. Ohuchi, S. Inaba, K. Okano, M. Fujiwara, T. Morooka, M. Goto, A. Kajita, T. Usui, K. Ishimaru, and Y. Toyoshima, "Guideline for low-temperature-operation technique to extend CMOS scaling," in *IEDM Tech. Dig.*, 2006, pp. 675–678.
[49] K. Azar, "Advanced cooling concepts and their challenges," presented at the Int. Workshop Thermal Investigations ICs and Systems (THERMINIC), Madrid, Spain, 2002. Invited Talk.
[50] P. Rodgers, V. Eveloy, and M. G. Pecht, "Limits of air-cooling: Status and challenges," in *Proc. SEMI-THERM*, 2005, pp. 116–124.
[51] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. EDL-2, no. 5, pp. 126–129, May 1981.
[52] R. S. Prasher, J.-Y. Chang, I. Sauciuc, S. Narasimhan, D. Chau, G. Chrysler, A. Myers, S. Prstic, and C. Hu, "Nano and micro technology-based next-generation package-level cooling solutions," *Intel Technol. J.*, vol. 9, no. 4, pp. 285–296, Nov. 2005. 4th quarter.
[53] R. Mahajan, R. Nair, V. Wakharkar, J. Swan, J. Tang, and G. Vandentop, "Emerging directions for packaging technologies," *Intel Technol. J.*, vol. 6, no. 2, pp. 62–75, May 2002. 2nd quarter.
[54] H. Y. Zhang, D. Pinjala, and P. S. Teo, "Thermal management of high power dissipation electronic packages: From air cooling to liquid cooling," in *Proc. Electron. Packag. Technol. Conf.*, 2003, pp. 325–620.
[55] C. LaBounty, A. Shakouri, and J. E. Bowers, "Design and characterization of thin film microcoolers," *J. Appl. Phys.*, vol. 89, no. 7, pp. 4059–4064, Apr. 2001.
[56] A. Shakouri and Y. Zhang, "On-chip solid-state cooling for integrated circuits using thin-film microrefrigerators," *IEEE Trans. Compon. Packag. Technol.*, vol. 28, no. 1, pp. 65–69, Mar. 2005.
[57] S. Ramanathan and G. Chrysler, "Solid-state refrigeration for cooling microprocessors," *IEEE Trans. Compon. Packag. Technol.*, vol. 29, no. 1, pp. 179–183, Mar. 2006.
[58] P. Wang and A. Bar-Cohen, "On-chip hot spot cooling using silicon thermoelectric microcoolers," *J. Appl. Phys.*, vol. 102, no. 3, 034503, Aug. 2007.
[59] K. Banerjee, S.-C. Lin, A. Keshavarzi, S. Narendra, and V. De, "A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management," in *IEDM Tech. Dig.*, 2003, pp. 887–890.
[60] P. E. Phelan, V. A. Chiriac, and T.-Y. Tom Lee, "Current and future miniature refrigeration cooling technologies for high power microelectronics," *IEEE Trans. Compon. Packag. Technol.*, vol. 25, no. 3, pp. 356–365, Sep. 2002.
[61] D. B. Go, S. V. Garimella, T. S. Fisher, and R. K. Mongia, "Ionic winds for locally enhanced cooling," *J. Appl. Phys.*, vol. 102, no. 5, 053302, Sep. 2007.
[62] S.-C. Lin, G. Chrysler, R. Mahajan, V. De, and K. Banerjee, "A self-consistent substrate thermal profile estimation technique for nanoscale ICs—Part I: Electrothermal couplings and full-chip package thermal model," *IEEE Trans. Electron Devices*, vol. 54, no. 12, pp. 3342–3350, Dec. 2007.
[63] S.-C. Lin, G. Chrysler, R. Mahajan, V. De, and K. Banerjee, "A self-consistent substrate thermal profile estimation technique for nanoscale ICs—Part II: Implementation and implications for power estimation and thermal management," *IEEE Trans. Electron Devices*, vol. 54, no. 12, pp. 3351–3360, Dec. 2007.
[64] J. Douglas and H. H. Rachford, "On the numerical solution of heat conduction problems in two or three space variables," *Trans. Amer. Math. Soc.*, vol. 82, no. 2, pp. 421–439, Jul. 1956.
[65] A. P. French and E. F. Taylor, *An Introduction to Quantum Physics*. New York: Norton, 1978.

Sheng-Chih Lin (S'03) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1996 and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 2007 under the tutelage of Prof. Kaustav Banerjee.

From 1998 to 2002, he was with the Phoenixtec Electronics Company, Ltd., and the CHROMA ATE Inc., respectively, in Taiwan. He joined Prof. Banerjee's research group at the University of California, Santa Barbara in Winter 2003. During the summer of 2005 and 2006, he worked as an intern in the Assembly and Test Technology Development of Intel in Chandler, Arizona. His research interests include electrothermal modeling and analysis of integrated circuits, variation-aware circuit design and optimization, and power/thermal management for nanoscale CMOS ICs. He has authored or coauthored over a dozen papers in journals and refereed international conferences.

Mr. Lin is a corecipient of the 2007 IEEE Micro Award.



Kaustav Banerjee (S'92–M'99–SM'03) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1999.

He was with Stanford University, Stanford, CA, from 1999 to 2001 as a Research Associate at the Center for Integrated Systems. From February to August 2002, he was a Visiting Faculty with the Circuit Research Laboratories, Intel, Hillsboro, OR. Since July 2002, he has been with the Faculty of the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. He has also held summer/visiting positions at Texas Instruments Incorporated, Dallas, TX, from 1993 to 1997, and the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2001. His research has been chronicled in over 140 journal and refereed international conference papers and in a book chapter. He has also coedited a book titled *Emerging Nanoelectronics: Life with and after CMOS* (Springer, 2004). His current research interests focus on nanometer-scale issues in high-performance/low-power very large scale integrated circuits (VLSI) as well as on circuits and systems issues in emerging nanoelectronics.

Dr. Banerjee has served on the technical program committees of several leading IEEE and ACM conferences, including IEDM, DAC, ICCAD, and IRPS. He has also served on the organizing committee of the International Symposium on Quality Electronic Design, at various positions including Technical Program Chair in 2002 and General Chair in 2005. Currently, he serves as a member of the Nanotechnology Committee of the IEEE Electron Devices Society. He has received a number of awards in recognition of his work, including the ACM SIGDA Outstanding New Faculty Award in 2004, a Research Award from the Electrostatic Discharge Association in 2005, a Best Paper Award at the Design Automation Conference in 2001, an Outstanding Student Paper Award at the VLSI/ULSI Multilevel Interconnection Conference in 2005, and an IEEE Micro Award in 2007.