8-1-2007

Towards a Thermal Moore's Law

Shankar Krishnan
Bell Laboratories

S V. Garimella
Purdue University, sureshg@purdue.edu

Gregory M. Chrysler
Intel Corportation

Ravi V. Mahajan
Intel Corporation

Krishnan, Shankar; Garimella, S V.; Chrysler, Gregory M.; and Mahajan, Ravi V., "Towards a Thermal Moore's Law" (2007). *CTRC Research Publications.* Paper 4.
http://dx.doi.org/10.1109/TADVP.2007.898517

Towards a Thermal Moore's Law

Shankar Krishnan[1], Suresh V. Garimella[1], Gregory M. Chrysler[2] and Ravi V. Mahajan[2]

[1]School of Mechanical Engineering
Purdue University
West Lafayette, IN 47907

[2]Intel Corporation
Assembly Technology Development
Chandler, AZ 85226

ABSTRACT

Thermal design power trends and power densities for present and future single-core microprocessors are investigated. These trends are derived based on Moore's law and scaling theory. Both active and stand-by power are discussed and accounted for in the calculations. A brief discussion of various leakage power components and their impact on the power density trends is provided. Two different lower limits of heat dissipation for irreversible logic computers that have previously appeared in the literature are discussed. These are based on the irreversibility of logic to represent one bit of information, and on the distribution of electrons to represent a bit. These limits are found to be two or more orders of magnitude lower than present-day microprocessor thermal design power trends. Further, these thermal demand trends are compared to the projected trends for the desktop product sector from the International Technology Roadmap for Semiconductors (ITRS) [1]. To evaluate the thermal impact of projected power densities, heat sink thermal resistances are calculated for a given technology target. Based on the heat sink thermal resistance trends, the evolution of a consistent air-cooling limit is predicted. One viable alternative to air-cooling, *i.e.*, the use of high-efficiency solid-state thermoelectric coolers (TECs), is explored in detail. The impact of different parasitics on the thermoelectric figure of merit (ZT) is quantified.

NOMENCLATURE

A	Area, cm^2
C	Capacitance, F
DF	Density factor, cm^{-2}
E	Energy, J
f	Frequency, Hz
I	Current, A
k_B	Boltzmann Constant, J/K
k	Thermal conductivity, W/mK
L	Channel length, m
l	TEC leg length, m
m	Swing parameter
n	Device parameter
N	Number of electrons

P	Power, W
Q	Heat transfer rate, W
q	Charge of an electron, eV
q”	Power density or heat flux, W/cm^2
R	Thermal impedance, $C\text{-}cm^2/W$
R_{el}	Electrical resistance, Ωm^2
S	Entropy, J/K
T	Temperature, K
V	Voltage, V
W	Channel Width, m
Z	Figure of merit, K^{-1}

Greek Symbols

α	Seebeck coefficient, V/K
α_{sf}	Switching activity factor
β	Scaling factor
ϕ	Coefficient of performance
γ	Linearized body effect coefficient
η	Drain induced barrier lowering coefficient or Packing coefficient
κ	Thermal conductance, W/K
μ	Carrier mobility
θ	Thermal resistance, C/W
ρ	Electrical resistivity, Ωm^2
σ	Electrical conductivity, $1/\Omega m^2$

Subscripts

a	Ambient
c	Configurational or Cold or Case or Core
d	Drain
F	Fermi
g	Gate
HS	Heat sink
h	Hot
i	Interstitial
j	Junction
l	Logical
n	N-type
ox	Oxide
p	P-type

r	Reference
s	Source or Heat sink
sub	Sub-threshold
s-c	Short-circuit
tec	Thermoelectric cooler
th	Threshold
0	Zero-bias

1. INTRODUCTION

The invention of the transistor in 1947 and the subsequent invention of integrated circuits in 1958 sparked the unprecedented growth of information processing technology employing silicon devices [2,3]. In the last five decades, the semiconductor industry has grown to become one of the largest industries in the world. The continued growth and development of the semiconductor industry has primarily relied on the exponential increase in the density of transistors in solid-state integrated circuits. The principal theme that has enabled such unprecedented growth is that decreasing the feature size of the unit transistor improves overall circuit performance while reducing the cost of the manufactured component. The performance of digital electronics has increased dramatically over the last five decades. Historically, the scaling of planar transistors has held to an observation by Gordon Moore in 1965 [4], known popularly as Moore's Law, which essentially forecasts that the number of transistors on a chip doubles every 18 months. While such an exponential trend cannot continue indefinitely due to fundamental limits imposed by basic physics, such barriers have to date been overcome, by innovations in materials processing techniques that have enabled faster and smaller devices and circuits [5]. Extensive literature exists on the difficulties of continued scaling and offers guidelines for future scaling of today's MOS technology [2,3,6,7,8].

The dissipation of energy in the form of heat has long been recognized as a potential issue that may limit information processing [4]. Both the supply of power and the removal of heat limit the performance of the information processing technology. In computing systems, information is represented by digitized electrical signals. In digital circuits, transistors are implemented as switches and valves. Electronic devices alter potential barriers that control the flow of carriers between power supplies and ground. In CMOS inverters, the most commonly used circuit family, currents flow during changes in logic states to charge the capacitances of devices and wires. Current flow ceases when a component voltage reaches the value of the voltage source. Energy is expended in charging and discharging the capacitances of the circuit, and the associated flow of carriers between power supply to ground generates heat. Additionally, in integrated circuits, many layers of interconnects generate heat due to Joule heating.

Increasing thermal management demand driven by increasing performance of semiconductor devices, typically represented by microprocessors, over the past two decades has led to the development of a number of innovative technologies [9,10,11,12]. As shown in [9], thermal demand is defined as the need to develop heat removal strategies that maintain the active surface of the microprocessor at or below a specific temperature while dissipating a highly non-uniform power. The non-uniformity of power across the chip surface tends to diminish the efficacy of thermal management solutions [9] and any assessment of heat removal solutions must accurately

comprehend the non-uniformity of power. Since on-chip power distributions tend to be proprietary information, and somewhat design-specific, the definition of strategic thermal demand becomes difficult. This can potentially lead to a non-optimal (both in terms of cost and performance) thermal solution and make setting strategic thermal directions for product families difficult. Thus the dependence on power distribution data must be reduced in setting strategic thermal directions. One path to achieving this is through the definition and use of a "density factor" to define efficacy of thermal management solutions [9]. Another is to attempt a definition of the strategic thermal demand; this is the focus of the work presented here.

The objectives of the present work are to forecast, based on the scaling theory for MOSFETs, the power densities that are expected to be dissipated by microprocessors and to analyze the capabilities of the various state-of-the-art cooling technologies to maintain microprocessor junction temperatures within acceptable limits. The specific objectives of the work are to forecast (1) the thermal design power trends and the associated thermal solution demands, (2) leakage power trends and their impact on cooling demand, (3) peak power density trends, and (4) the limits of air cooling.

Further, based on the thermal solution demands, an analysis is performed on the impact of different parasitics on the performance of Thermoelectric Cooler (TEC) modules, a field that has attracted recent interest [13].

2. POWER DISSIPATION

In this section, the fundamental (lower) limits on the power dissipation are discussed followed by a brief discussion of power dissipation by transistors [6]. Since the objective of the present work is to project the power dissipation trends of microprocessors, a brief overview of the scaling theory employed to improve CMOS technologies along a Moore's Law trajectory is discussed. A detailed description of leakage power and its impact on thermal management is then presented.

2.1 Lower Limits on Power Dissipation for Irreversible Logic Computing

According to a basic principle of the thermodynamics of computations, *i.e.*, the Landauer principle [14], any irreversible manipulation of information, such as erasure of a bit, which throws away the computer's previous logical state must be accompanied by an increase in entropy in the processing environment. Erasure of a bit changes the logical entropy of the system from $k_B \log 2$ to $k_B \log 1$ (= 0), where k_B is the Boltzmann constant. Then the bit erasure essentially is associated with an energy dissipation of $k_B T \log 2$ per bit. This is also referred to as the Shannon-von Neumann-Landauer limit and is the minimum energy that should be supplied for any computation via irreversible logic; it is expressed as

$$E_l = (k_B \log 2) T \tag{1}$$

This dissipation of energy is only due to the act of erasure and is independent of any other internal energy and entropy associated with the physical representation of the bit [15]. It is also known that the same amount of energy, E_l, is dissipated during signal transmission by interconnects [3, 6,16].

In addition to logical entropy, entropy can also be associated with the distribution of electrons making up a bit. Using the Sommerfeld approximation, valid for temperatures that are low compared to the Fermi temperature, the electronic entropy per bit of information is defined as [15]

$$S_c = \frac{N\pi^2 k_B T}{2T_F} \tag{2}$$

where N is the number of electrons (= C_{load}/q) per bit and T_F is the Fermi temperature (~ 10^5 K). The temperature of the bit is assumed to be the same as that of the chip. The configurational energy consumed for processing one bit of information (if the bit was cooled) is given by

$$E_c = TdS_c \tag{3}$$

Equations (1) and (3) provide the lower bound on energy consumption by irreversible logic computers; however, the practical applicability of these equations must be called into question since the energy consumption predicted by equations (1) and (3) is three or more orders of magnitude lower than the electrostatic energy consumed by present-day electronics.

2.2 Power Estimation for Typical Microprocessors

In digital circuits, transistors are implemented as switches. Current flow between the source and drain terminals in a transistor is controlled by the voltage at the gate terminal. The gate is electrically isolated from the silicon by a thin insulating layer. The gate influences the device through the electric field resulting from different gate biases. A schematic of a MOS transistor is shown in Figure 1.

In general, for CMOS circuits, power is dissipated during standby and active modes. The total power dissipation during the active mode consists of static and dynamic components. Switching power or dynamic power dissipation is due to charging and discharging of capacitive load ($\alpha_{sf}CV^2 f_{clock}$). Dynamic power consumption occurs whenever there is a power-consuming voltage transition in the device. The switching activity factor (α_{sf}) is defined as the effective number of power-consuming voltage transitions (*e.g.*, the $0 \rightarrow 1$ transition is a power-consuming transition) experienced by the output capacitive load. The switching activity factor depends on the circuit topology and logic style, as well as on the input signal statistics. For a two-input logic gate where the inputs are uniform and equally probable, the switching factor can be expressed as P0 × P1 where P0 and P1 are the probabilities that 0 or 1 appear as the output. Therefore the activity factor can be defined as the mean probability of the possible outputs. The evaluation of switching activity factors for large circuits is complicated as the input signals are not uniform. Switching activity is one of the main reasons for the higher power consumption of logic transistors where the activity factors are very high compared to cache transistors. A smaller component of dynamic power dissipation occurs due to the short-circuit current, also called crow-bar current, that flows momentarily when both PMOS and NMOS in an inverter are simultaneously conducting during a state change.

Static power dissipation is due to the various leakage currents arising due to device limitations in the transistor. A schematic of the dynamic and static power consumption mechanisms in MOSFETs is shown in Figure 2. The total power dissipation during the active mode can be expressed as

$$P_{tot} = \alpha_{sf}\left(C_{load}V_{dd}^2\right) f_{clock} + I_{s-c} \cdot V_{dd} + I_{leak} \cdot V_{dd} \tag{4}$$

The first two terms on the right hand side are components of the dynamic power and the third term is the static (or leakage) component. In equation (4), α_{sf} is the activity factor, C_{load} is the load capacitance that includes gate, junction and interconnect (or line) capacitances, V_{dd} is the supply voltage, f_{clock} is the clock frequency, $I_{s\text{-}c}$ is short-circuit (or crow-bar) current and I_{leak} is leakage current. The switching power has been the most dominant component of power dissipation in circuits. But the leakage power is increasing at an exponential rate mainly due to sub-threshold conduction and gate-oxide tunneling [5]. A brief discussion of leakage power is presented later in this paper. It should be noted that equation (4) applies to an isolated transistor. If the distribution of transistor geometries is assumed to be the same across large groups of transistors implemented on the silicon, then this equation can be generalized for the entire chip by multiplying with the average number of transistors. However, the size of transistors varies significantly on the die. In view of the assumption of uniform transistor size, equation (4) only provides an order of magnitude estimation. In general, the area of the memory cache transistors is smaller than the logic transistors and the transistor density in the cache is higher than the logic transistor density.

2.3 Scaling Theory

Historically, the total numbers of transistors on a chip have doubled every 12 months [4]. By the early '80s the scaling of transistor sizes, and hence the number of transistors, slowed down due to increased complexities involved in printing the device dimensions onto the chip. The total number of transistors in the '90s doubled every 18 to 24 months, bearing out Moore's Law. This period of 18 to 24 months is commonly referred to as a technology generation; the microprocessor technology for that generation is characterized by the minimum feature size (gate length), which is known as a technology node [1].

Scaling of semiconductor technology increases the transistor density, improves the transistor performance and reduces transistor power consumption. Three defined goals for current technology scaling [17] are to: (1) reduce gate delay (~ CV/I) by 30% and thus increase the operating frequency (f ~ I/CV) by 43% [1/(1-0.3)]; (2) double the transistor density; and (3) reduce the energy per transition (CV^2) by 65% and hence, the power consumption (CV^2f) by 50%, where C and V scale as 0.7 per generation along with linear dimensions to follow Moore's law. These technological improvements are achieved through advancements in architecture, circuit design and fabrication technology. Over the last two decades, two different scaling methods have mainly been employed to achieve these technological improvements.

The two scaling approaches are: (1) constant voltage scaling where the supply voltage is maintained constant, and (2) constant electric field scaling where vertical and horizontal electric fields are maintained constant. If supply voltage remains constant (constant voltage scaling), and if the device dimensions are shrunk by 30%, the capacitance scales down by 30% ($C \sim \beta$) and the frequency increases by 43%, since $f \sim I/CV$ $(1/\beta)$ where I also scales by a factor of β. As result, the power remains constant: $P \sim CV^2f \rightarrow 0.7 \times 1^2 \times (1/0.7) = 1$. A schematic diagram depicting constant voltage scaling is shown in Figure 3. Present-day technologies employ constant electric field scaling where the device dimensions and supply voltages are scaled. Scaling of the supply voltages reduces the switching power dissipation (quadratic influence). Constant electric field scaling is illustrated in Figure 4.

However, scaling the supply voltage leads to a severe penalty in speed of operation of the circuits. This is because the propagation delay (time required for signal to travel from input to output) increases as the voltage is scaled:

$$\tau \propto CV/(V-V_{th})^n \quad (5)$$

This increase in propagation delay limits the circuit speed and hence the switching frequency. In order to lower the power consumption without significant performance loss, the electric field scaling methodology also employs threshold voltage scaling [17,18]. It should be noted that the supply and the threshold voltages are not scaled by the same factor. This scaling requires a decrease in threshold voltage which increases the static power dissipation in the current technologies, *i.e.*, it increases the off-state current I_{leak}. In constant field scaling, the device dimensions and voltage are scaled by 30%, the capacitance scales by 30% and frequency increases by 43%. Therefore, the power P scales by the factor $\sim 0.7 \times 0.7^2 \times (1/0.7) = 0.5$. A detailed description of the effect of constant field scaling on various device parameters such as capacitance, resistance, and gate delay is available in [17].

Figure 5 shows the maximum power dissipation of several Intel microprocessors, up to the 0.13 μm node. Technology nodes prior to 0.8 μm employed constant voltage scaling and hence the power increased dramatically due to doubling of transistors (though power consumed by a single transistor remained constant) every generation. Constant field scaling has been used thereafter. It is seen that when the microprocessors move to the next generation with constant voltage scaling, the power remains constant, as against constant field scaling where the power decreases when the technology is improved. This is consistent with the scaling arguments discussed above. The leakage or static power in technology nodes before 0.18 μm was less than 5% of the total power. It should be noted that the power dissipation is not governed just by technology scaling, but also by the micro-architecture implementation, circuit design, and other parameters.

2.4 Leakage Power: Mechanisms and Trends

In this section, three important leakage mechanisms in current technologies are briefly discussed. A detailed review of leakage mechanisms and their reduction methodologies is available in [19].

2.4.1 pn Junction Reverse-Bias Current

Figure 6 schematically illustrates various leakage mechanisms in a MOS transistor. Typically, the source and drain to well junctions are reverse-biased. Junction leakage occurs due to minority carrier diffusion/drift near the edge of the depletion region and electron-hole pair generation in the depletion region of the reverse-biased junction. Also, if both n and p regions are heavily doped (high electric field) to attain the desired threshold voltage, as is the case in current technologies, band-to-band tunneling occurs, *i.e.*, electrons from the valence band of the p region tunnel through to the conduction band of the n region. In such situations, band-to-band tunneling dominates the pn junction leakage current.

2.4.2 Sub-threshold Leakage Current

The gate delay of a transistor (the approximate time required to charge the gate capacitance) can be expressed as $\tau = C_{gate} \cdot V_{dd}/I_{d,sat}$, where C_{gate} is the gate capacitance per unit width, V_{dd} is the supply voltage and $I_{d,sat}$ is the saturation drain current per unit width of the transistor. Capacitance C_{gate} is proportional to the ratio of channel length to oxide thickness. As the device dimensions shrink by a factor k, C_{gate} remains constant. But the

drain current is proportional to $(V_{dd} - V_{th})^n$, where n ranges between 1 to 2. As V_{dd} is scaled by k, in order to maintain the drain current I_d, V_{th} should also be reduced. This reduction in V_{th} along with the device dimensions causes an increase in sub-threshold leakage. Sub-threshold leakage current occurs when the gate voltage is below the threshold voltage and mainly consists of diffusion current. When the gate voltage is less than the threshold voltage, the number of mobile carriers and the longitudinal electric field are small, and hence the drift component of the sub-threshold drain-to-source current is negligible. Therefore, sub-threshold conduction is dominated by the diffusion current, and carriers move by diffusion along the surface below the channel. The sub-threshold leakage of a MOS device can be expressed as [19,20]

$$I_{sub,leak} = A_0 \exp\left(\frac{\left(V_g - V_s - V_{th0} - \gamma V_s + \eta V_{ds}\right)}{m\left(k_B T/q\right)}\right)\left(1 - \exp\left(\frac{-V_{ds} q}{k_B T}\right)\right) \tag{6}$$

where $A_0 = \mu_0 C_{ox} \frac{W}{L}(m-1)\left(\frac{k_B T}{q}\right)^2 \exp(1.8)\exp\left(-\frac{\Delta V_{th} q}{\eta k_B T}\right)$, μ_0 is the zero-bias mobility, C_{ox} is the gate oxide capacitance, W and L are the channel width and length, m is the sub-threshold swing coefficient (body effect coefficient), T is the temperature, q is the electron-charge, η is the drain induced barrier lowering (DIBL) coefficient, γ is the linearized body effect coefficient, ΔV_{th} is a parameter accounting for the transistor-to-transistor leakage variation, V_{th0} is the zero bias threshold voltage, and V is the voltage. It should be noted that the sub-threshold leakage current depends exponentially on threshold voltage and temperature, $I_{sub,leak} \propto \exp(-V_{th}/T)$, *i.e.*, any further decrease from the present-day values of threshold voltage and increase in transistor temperature would increase the leakage current and static power exponentially. The temperature dependency of sub-threshold leakage power is brought out in Figure 7. The sub-threshold leakage power is calculated using the expression [20]

$$I_{sub,leak} = B \cdot \exp\left(\frac{-qV_{th}}{nk_B T}\right) \tag{7}$$

Here, B and n are device parameters and are usually obtained from experiments. In Figure 7, a threshold voltage of 0.34 V and n = 1.4 [21] are used. In Figure 7, the vertical axis is the sub-threshold leakage power normalized with the peak temperature of 100°C (ratio of leakage power at any given temperature to leakage power at 100°C) and the horizontal axis is the temperature. As the maximum allowable temperature of the device is reduced from 100°C to 70°C, the sub-threshold leakage power reduces by 50%. Significant power reduction can be obtained if the junction temperature is reduced. This dependence on temperature of leakage power (and hence, of total thermal design power) has serious implications in microprocessor testing and noise immunity [22].

2.4.3 Gate-Oxide Tunneling

As the device dimensions shrink, the gate-oxide thickness also reduces. This reduction in gate-oxide thickness causes an increase in the electric field across the oxide. The high electric field coupled with low oxide thickness results in tunneling of electrons. The tunneling of electrons between the p-type substrate and gate can be either Fowler-Nordheim (FN) tunneling where the electrons tunnel into the conduction band of the oxide layer, or direct tunneling where the electrons directly tunnel to the gate through the forbidden energy gap of the SiO_2 layer [19]. FN tunneling occurs only when the oxide layer is thick (as in earlier technology generations) and electric field

across the layer is high. In very thin oxide layers (< 3 nm), electrons from the silicon surface directly tunnel to the polysilicon gate. It has been reported that the minimum allowable gate-oxide thickness for MOS transistors is 0.7 nm; below this thickness, the gate oxide breaks down [23]. The suggested alternative solution to SiO_2 is to use a high dielectric constant material for the gate electrode [23]. Other leakage mechanisms such as punch-through and gate-induced drain leakage and techniques for leakage reduction are detailed in [19].

The two major leakage currents plaguing present-day microprocessors are gate-oxide tunneling and sub-threshold leakage current. At the device level and at room temperature, the contributions of gate-oxide tunneling current and sub-threshold leakage current are similar in relative magnitude. But as microprocessors are operated at higher temperatures (~ 373 K), the sub-threshold leakage dominates the gate-oxide tunneling. In general, leakage is highest in "hot spots" because the temperature is high in these regions leading to higher sub-threshold leakage. In the SRAM memory cache, gate-oxide tunneling current dominates sub-threshold current, while in core logic the converse is observed. A detailed account of contributions from various leakage currents to the total leakage power is beyond the scope of this work.

The temperature dependency of the leakage power presents significant challenges to the thermal management of microprocessors. Figure 8 illustrates the typical junction temperature vs. thermal resistance plot [24] used to define the cooling solution (heat sink) requirements. Thermal resistance (θ) is the resistance to heat flow from one point to another (*e.g.*, θ_{ja} is resistance to heat flow from the junction to ambient) and is the ratio of temperature difference between the points to the power dissipated, $\theta = \Delta T / P$. In switching-power-dominant technologies where leakage power is negligible, the junction temperature vs. thermal resistance plot is a straight line (solid black line in the figure). If the junction temperature is increased, the requirements on the heat sink are relaxed. In leakage-power-dominant technologies on the other hand, shown as a red dashed line in Figure 8, an increase in junction temperature aggravates the situation by increasing the power dissipation requirements and hence decreasing the allowing thermal resistance. Significant power savings can be extracted when microprocessors are operated below present-day device operating temperatures.

2.5 Estimated Power Trends

Figure 9 shows the estimated average power trends for single core micro-architecture. Plotted in Figure 9a are the lower bounds on the power density trends calculated from equations (1) and (3). In calculating the power density from equations (1) and (3) it was assumed that all the transistors are active. Figure 9 also show an example estimated trend based on equation (4). In the projected power density curve, the leakage power was fixed at 30% of the total power after 2005, while the short-circuit power was fixed at 5% of the switching power. The initial values and expected trends for the various parameters used to calculate the trends are listed in Table 1 [25,26,27,28]. The voltage was assumed to scale by 15% every technology generation [25]. The frequency scaling was deduced from Intel datasheets provided in [26]. A typical switching activity factor (α) value of 0.1 was used for all the calculations [27,28]. The die size is kept constant at 1.4 cm^2. Also plotted in Figure 9 are published data from ITRS 2004 [1]. Based on the ratio of logic to memory power density (~ 10) [29], the hot spots can be estimated, roughly, to be 3 to 8 times the average power density. The peak flux range is established by considering the cache transistors to constitute between 50% and 90% of the total (cache + logic) transistors on the chip, and taking a geometric mean

of their contribution to average power density ($q_{cache}^{\%cache} \times q_{logic}^{(1-\%cache)}$). Thus, with 50% cache and 50% logic, and a logic to cache power density ratio of 10, the average flux is 3.16 (= $1^{0.5} \times 10^{0.5}$). Therefore, the ratio of hot-spot power density (only logic) to average power density is 10/3.16 (~ 3).

Estimates such as those developed here for average and peak power density offer a mechanism to determine the thermal impact of the power density projections, and to develop commensurate thermal solutions. However they are not intended to provide projections on future directions since these estimates do not account for the impact of low-power circuit techniques and architectural innovations such as multi-core and multi-thread processing which have a significant impact on the evolution of thermal demand. These projections should be assumed to bound only single-core and single-thread application-based power density evolutions.

3. THERMAL MANAGEMENT OF MICROPROCESSORS

The thermal impact of the projected power density trends of microprocessors in terms of heat sink requirements is now discussed. Cooling solution demands are typically expressed in terms of the allowable thermal resistance to heat flow from junction (transistor) to ambient as a function of thermal design power. The definition of thermal resistance is general and holds for different thermal architectures of which two types may be considered: one, where a thermal interface material (TIM) is used to thermally couple a bare die to a heat sink, and a second, where the heat sink interfaces with an integrated heat spreader (IHS) through a TIM [9]. For non-uniform power distribution on the die, a simple metric called the density factor for characterizing the extent of non-uniformity is suggested in [9]. For non-uniform on-die power density distribution, the thermal design power (TDP) can be expressed as

$$TDP = \frac{\left(T_j - T_a\right)}{\left(DF_{jc} \cdot R_{jc} + \theta_{cs} + \theta_{sa}\right)} \tag{8}$$

In the above equation, the subscripts j,c,s and a refer to junction, case, sink and ambient, respectively. In equation (8), R is the thermal impedance defined as the thermal resistance normalized by die area for a uniformly powered die, and is expressed as $R_{jx} = \frac{\left(T_{j,\max} - T_x\right)}{TDP} \times A_{die}$. In equation (8), DF is the density factor which accounts for non-uniformity in the power distribution and is defined as the ratio of thermal resistance to thermal impedance. The density factor can be mathematically expressed as

$$DF_{js} = \frac{\theta_{js}}{R_{js}} \tag{9}$$

The density factor quantifies the impact of die size and die-power maps on the thermal resistance from junction to any specific location along the primary heat flow path.

In Figure 10 the evolution of heat sink requirements ($\theta_{ca} = \theta_{cs} + \theta_{sa}$) is plotted based on the projected average power density (q_{avg}") calculated above for various technology generations. The heat sink resistance requirement, θ_{ca}, can be obtained using the following equation

$$(\theta_{ca})_{required} = \frac{\left(\left(T_{j,\max} - T_a\right) - \left(T_{j,\max} - T_c\right)\right)}{TDP} \tag{10}$$

where $(T_{j,max} - T_c) = R_{jc} \times q_{avg}$". A package thermal impedance value R_{jc} of 0.1-0.3 °C-cm^2/W was used along with junction to ambient temperature difference ($T_{j,max} - T_a$) of 60°C to calculate the heat sink resistance. The solid lines and symbols represent the case for a uniformly powered die (DF = $1/A_{die}$ from equation 9). The dashed lines and hollow symbols show the impact of power non-uniformity (DF = 1.5 cm^{-2}) on the required second-level thermal package in order to maintain the junction to ambient temperature difference of 60°C.

Refrigeration cooling, in principle, can provide negative thermal resistance. Solid-state thermoelectric refrigeration has recently been shown to have the capability of cooling very high power densities [13]. Due to the high heat flux handling potential of solid-state thermoelectric refrigeration, and since this technology has the potential of scaling along with circuit technology, it is explored in the next section as one of the viable options for cooling future microprocessors. There are, of course, many other viable cooling options in the literature, such as single or two-phase liquid cooling. These are not discussed here and solid-state refrigeration is considered as an illustrative example. A number of reviews of liquid cooling and other viable alternative technologies can be found in [30,31,32].

4. SOLID-STATE REFRIGERATION

The potential benefits of using system-level refrigeration cooling of microprocessors have been reported in previous studies such as [33]. These benefits include higher carrier mobility and thus carrier saturation velocity (higher speed), lower interconnect resistances (higher speed), reduced sub-threshold currents (lower power), lower electro-migration and thermal wear out (higher reliability) [34]. Vassighi et al. [35] studied the trade-offs in microprocessor frequency and system power achievable for low-temperature operation using refrigeration for the 130 nm technology node. For the case where the power dissipated by the microprocessor with and without refrigeration was maintained constant, an 11% increase in frequency was reported in the presence of refrigeration. For the case where the frequency was constant, a 38% reduction in power (compared to air cooling) was reported.

4.1 Background

Solid-state thermoelectric cooling has the advantages of long life, no moving parts, no emissions of toxic gases, low maintenance, and high reliability [36]. Another major advantage of solid-state refrigeration is that it can be structured into modular units and can be miniaturized. Solid-state thermoelectric refrigerators utilize the Peltier effect to pump heat from one location to another. When current is passed through two dissimilar metals or metal-semiconductor pairs that are connected to each other at two junctions, heat is released at one junction and absorbed at the other [37]. Other refrigeration technologies that use the Peltier effect to pump heat are solid-state thermionic and thermo-tunneling (thermionic) refrigerators.

Table 2 compares the various thin-film refrigeration technologies which exploit the Peltier effect [38,39,40,41]. In solid-state thermionic coolers, an emitter and collector consisting of either metal or a heavily doped semiconductor are separated by a solid-state barrier material which is made of a less heavily doped

semiconductor. The barrier is made thin enough so that electrons travel ballistically through the barrier layer. Due to the Peltier effect at the barrier/metal (or semiconductor) junction, electrons pass through the barrier layer from the emitter to the collector causing cooling at the emitter-barrier junction and heating at the barrier-collector junction [38]. The primary difference between solid-state thermionic and thermoelectric devices arises because of the ballistic transport of electrons through the barrier region in the former. Compared to thermoelectric devices, the electrical current through a thermionic device is nonlinear and there is no Joule heating in the barrier layer for an ideal thermionic device. A vacuum thermionic device has two parallel plates separated by vacuum. This technology can be considered as an extension of thermionic cooling that combines the effect of thermionic cooling (evaporative emission) with quantum mechanical electron tunneling [39]. There are many variations possible to thermionic refrigerators depending on the design of the barrier. Though thermo-tunneling seems to have high theoretical efficiency compared to solid-state refrigerators, solid-state thermoelectric refrigerators have the advantage of having been studied in detail by many previous investigators and their capabilities and limitations are well known [40,41]. Also, recent developments in both bulk and low-dimensional TE materials (the latter exploiting quantum confinement effects) have improved the thermoelectric figure of merit significantly [42]. Hence the present study focuses on solid-state thermoelectric refrigerators.

Solid state cooling based on the Peltier effect was discovered as early as 1800 [42]. Extensive technical literature exists on solid-state thermoelectric energy conversion devices [42,43]; a detailed review of the literature is beyond the scope of this work. A typical thermoelectric cooler is shown in Figure 11. The p-type and n-type semiconductor elements are connected on the hot and cold sides such that a current flows through all the elements in series (if they were in parallel, the effects would cancel out). Practical devices are made of many pairs of p-n legs, usually arranged such that current flows in series through the legs and energy flows in parallel from the cold side to the hot side. Efficient solid-state refrigerators based on the Peltier effect require materials of high electrical conductivity (σ), high Seebeck coefficient (α) and low thermal conductivity (k) [42,44]. All these requirements can be combined into a figure of merit for thermoelectric energy conversion:

$$Z = \frac{\alpha^2 \sigma}{k} \tag{11}$$

The higher the value of Z for a thermoelectric material, the higher is its efficiency. The recent resurgence in thermoelectrics research inspired by developments in nano-structured and bulk materials is leading to the achievement of higher Z values [13].

4.2 Mathematical Modeling

For an ideal thermoelectric device (Figure 12), under steady-state conditions, the heat absorbed at the cold junction and rejected at the hot junction can be written as [42,45]

$$Q_c = \alpha I T_c - \kappa \left(T_h - T_c\right) - \frac{1}{2} I^2 R \tag{12}$$

$$Q_h = \alpha I T_h - \kappa \left(T_h - T_c\right) - \frac{1}{2} I^2 R \tag{13}$$

where $\kappa = \frac{A}{l}\left(k_p + k_n\right)$ and $R = \frac{l}{A}\left(\rho_p + \rho_n\right)$. In equations (12) and (13), α is the Seebeck coefficient, I the current, k the thermal conductivity, ρ the electrical resistivity, and l and A are the length and cross-section area of the thermoelectric element. In order to pump heat from the heat source, the external power that should be supplied is given by

$$P_{tec} = Q_h - Q_c = I^2 R + \alpha \Delta T \tag{14}$$

It should be noted that equations (12) and (13) do not account for parasitic losses such as interface losses due to contact resistances, back conduction through the interstitial material and other parasitic losses. In equations (12) and (13), the first term on the right hand side accounts for the Peltier effect, the second term accounts for heat conduction and the third accounts for Joule heating as current passes through the thermoelectric material.

The following (parasitic) losses, besides the intrinsic losses due to Joule heating and heat conduction through the TEC leg, should be accounted for in the design of a TEC module:

- electrical contact resistance offered by the inclusion of anti-diffusion layers and solders [46,47],
- thermal contact resistance due to electron and phonon boundary resistances at the thermal/metal interfaces, and non-equilibrium effects (*i.e.*, interaction between phonons and electrons) [48],
- substrate thermal resistance and spreading resistance at the hot and cold sides [49], and
- back heat conduction through the interstitial materials and supporting frames.

The effects of electrical contact resistance and back conduction through the interstitial material on the effective Z of the material are discussed further in the following as these are the major parasitic losses that limit the miniaturization of Peltier coolers.

4.3 Electrical Contact Resistance

Figure 13 shows a thermoelectric cooling leg with various sources of electrical resistance [50]. The representation in Figure 13 shows the various layers of material used besides the metal electrode and thermoelectric material. All these layers contribute to the electrical contact resistance. As current is passed through these layers, Joule heating at the interface where the Peltier effect occurs lowers the effectiveness of the thermoelectric modules [50]. The definition of Z for a practical thermoelectric device can be recast as follows [51, 52]

$$Z' = \frac{\alpha^2}{\kappa R} = \frac{\alpha^2}{\rho k}\left[\frac{1}{1 + 2R_{el}/\rho l}\right] \Rightarrow \frac{Z'}{Z} = \left[\frac{1}{1 + 2R_{el}/\rho l}\right] \tag{15}$$

in which R_{el} is the electrical contact resistance (Ωm^2). In equation (15) Z' is the modified or effective figure of merit of the thermoelectric device.

Figure 14 shows the effect of electrical contact resistance on the figure of merit as a function of leg length of the thermoelectric device. Two bounds on R_{el} are used, the upper bound being the microfabrication limit on the order of 10^{-11} Ωm^2 [53] and the assumed lower-bound worst case R_{el} of 10^{-8} Ωm^2. It can be observed that R_{el} poses constraints on the miniaturization as the contact resistance itself is enough to bring down the effective Z' to less than 10% of its intrinsic value (Z). From equation (15), it is seen that R_{el} should be much lower than the product of leg

length and electrical resistivity of the TEC material ($R_{el} << \rho l$) for a high operating figure of merit (Z'). As discussed earlier, the shorter the leg length of the TEC, the higher is its cooling power density ($q_c" \propto l$). The effect of electrical contact resistance on the heat dissipation capacity of the thermoelectric device can be easily accommodated in equations (12)-(14) by lumping them into the resistance term (R). Other interface effects on Z are discussed in detail in [53].

4.4 Back Heat Conduction

As heat is pumped from the cold junction to the hot junction, there is back flow of heat from the hot to the cold junction through the thermoelectric element and also through the interstitial material between the thermoelectric materials. A schematic of the back conduction in a typical thermoelectric element is depicted in Figure 15. The packing density (η) of the thermoelectric material may be defined as the ratio of the area of the thermoelectric material to the total area of the thermoelectric module, $\eta = A_{tec}/(A_{tec} + A_i)$. The effect of back conduction through the interstitial material on the operating Z' of the TEC can be expressed as

$$Z' = \frac{\alpha^2}{KR} = \frac{\alpha^2}{\rho k_{tec}} \left[\frac{1}{1 + \frac{k_i}{k_{tec}} \left(\frac{1}{\eta} - 1 \right)} \right] \Rightarrow \frac{Z'}{Z} = \left[\frac{\eta}{\eta k_{tec} + (1-\eta) k_i} \right] \tag{16}$$

Using equation (16), the effect of back conduction is illustrated in Figure 16. It may be noted that equation (16) reduces to the appropriate limit for $\eta = 0$ (Z = 0 because of no thermoelectric material) and $\eta = 1$ (Z' = Z). It is also observed that a higher packing density and lower k_i/k_{tec} ratio result in a higher operating figure of merit. The effect of back conduction of heat through the interstitial material in the TEC module can be included in the mathematical model (equations 12-14) by using the unit-cell model and lumping the interstitial thermal conductivity (k_i) and packing density (η) in κ as

$$\kappa = \frac{A}{l} k_{tec} \left(1 + \left(\frac{1}{\eta} - 1 \right) \frac{k_i}{k_{tec}} \right) \tag{ }$$

For achieving a high operating Z' for the TEC module, the electrical contact resistance and the back conduction through the interstitial material should be reduced. The electrical contact resistance can be lowered by using microfabrication techniques to prepare the metal TE leg junction. The back conduction through the interstitial material can be reduced by using highly insulating materials and through higher packing density.

4.5 Impact of Loss Mechanisms on TEC Efficiency

The two loss mechanisms described in the previous two sections not only reduce the effective ZT value, but they also decrease the efficiency of the TEC module. The electrical contact resistance increases the internal heating (I^2R) of the TEC itself and hence reduces the amount of energy it can pump. Back conduction also reduces the net heat pumping. The coefficient of performance (COP) is the typical efficiency metric for refrigeration systems and is defined as the ratio of energy pumped to the energy input required to achieve the pumping ($COP = Q_c/(Q_h-Q_c)$). TECs can be operated under many different conditions such as maximum heat pumping ability, maximum COP, and

optimal COP and heat pumping conditions depending on the driving current (usually normalized with the optimal current, i.e., I/I_{opt}). It is well known that the maximum COP is higher for higher ZT material [42]. A detailed discussion on the effect of ZT on the COP can be obtained from references [43,45,53,54]. Hence, it is readily apparent that the loss mechanisms would decrease the maximum COP and move the point of this reduced maximum to a different current. The movement, however, is dependent on the loss mechanism. Increasing the electrical contact resistance pushes the maximum COP to a lower value of I/I_{opt}. This is due to the increased self-heating caused by the additional resistance. The addition of the back conduction, on the other hand, pushes the maximum COP to a higher value of I/I_{opt}, since it now takes additional heat pumping, from the driving current, to compensate for the back conduction of heat through the material surrounding the TEC elements. While the changes described are qualitatively independent of the scale of the TEC module, the impact on the magnitude of heat pumping is strongly dependent on both the values of the loss mechanism and the scale of the TEC module.

5. CONCLUSIONS

A study was undertaken to estimate the thermal design power and peak power density trends of single-core microprocessors. Two different lower bounds on the power density are discussed and they are found to be two orders of magnitude less than the present-day power densities. A review of various components of power dissipated by microprocessors is provided and the impacts of leakage power on the cooling solutions are discussed. Average and peak power density trends are estimated based on technological scaling trends for single-core micro-architectures. Heat sink thermal resistance requirements for projected power densities are also calculated. Projections from this work show that significant improvements in packages and heat sinks for system cooling continue to be required.

Solid-state thermoelectric refrigeration is a promising technology for high power density cooling and so its limitations are studied. Thin-film thermoelectric coolers (TECs) are particularly attractive for cooling high heat fluxes, but thinning of a TEC, while yielding a higher cooling capability, also results in greater parasitic losses.

Acknowledgements

The authors gratefully acknowledge discussions with David Chau, Ross Farahani, Rajashree Baskaran, Chuan Hu, Ioan Sauciuc, Siva Narendra, Ted Burton, Shekhar Borkar, Vivek De and Ali Keshavarzi of Intel Corporation.

6. REFERENCES

[1] Semiconductor Industry Association (SIA), International Technology Roadmap for Semiconductors, 2004.
[2] R. W. Keyes, "Fundamental limits of silicon technology," *Proceedings of the IEEE*, Vol. 89, pp. 227 – 239, 2001.
[3] J. D. Meindl, "Low power microelectronics: Retrospect and prospect," *Proceedings of the IEEE*, Vol. 83, pp. 619 – 635, 1995.
[4] G. E. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, Vol. 86, pp. 82 – 85, 1998.

[5] G. E. Moore, "No exponential is forever: But "forever" can be delayed!," *IEEE International Solid-State Circuits Conference*, 2003.
[6] V. V. Zhirnov, R. K. Calvin, J. A. Hutchby, and G. I. Bourianoff, "Limits to binary logic switch scaling – A Gedankan model," *Proceedings of the IEEE*, Vol. 91, pp. 1934 – 1939, 2003.
[7] L. B. Kish, "End of Moore's law: thermal (noise) death of integration in micro and nano electronics," *Physics Letters A*, Vol. 305, pp. 144 – 149, 2002.
[8] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur and H. P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proceedings of the IEEE*, Vol. 89, pp. 259 – 288, 2001.
[9] J. Torresola, C. Chiu, G. Chrysler, D. Grannes, R. Mahajan, R. Prasher and A. Watwe, "Density factor approach impact on die power maps on thermal management," *IEEE Transactions on Advanced Packaging*, Vol. 28, pp. 659 – 664, 2005.
[10] S. V. Garimella, "Advances in mesoscale thermal management technologies for microelectronics," *Microelectronics Journal* (available online), 2005.
[11] S. V. Garimella, "Transport in mesoscale cooling systems," *Symposium on Fundamental Issues and Perspectives in Fluid Mechanics*, FEDSM2005-77325, ASME Fluids Engineering Summer Conference, Houston, TX, June 19-23, 2005.
[12] C. J. M. Lasance and R. E. Simons, "Advances in high-performance cooling," *Electronics Cooling*, Vol. 11, pp. 22 – 39, 2005.
[13] R. Venkatasubramanian, E. Siivola, T. Colpitts and B. O'Quinn, "Thin-film thermoelectric devices with high room-temperature figures of merit," *Nature*, Vol. 413, pp. 597 – 602, 2001.
[14] C. H. Bennett, "Notes on Landauer's principle, reversible computation, and Maxwell's demon," *Studies in History and Philosophy of Modern Physics*, Vol. 34, pp. 501-510, 2003.
[15] N. Gershenfeld, "Signal entropy and the thermodynamics of computation," *IBM Systems Journal*, Vol. 35, pp. 577 – 586, 1996.
[16] J. D. Meindl and J. A. Davis, "The fundamental limit on binary switching energy for terascale integration (TSI)," *IEEE Journal of Solid-State Circuits*, Vol. 35, pp. 1515 – 1516, 2000.
[17] S. Borkar, "Design Challenges of technology scaling," *IEEE Micro*, Vol. 19, pp. 23 – 29, 1999.
[18] Z. Chen, J. Shott, J. Burr, and J. D. Plummer, "CMOS technology scaling for low voltage low power applications," *IEEE Symposium on Low Power Electronics*, pp. 56 – 57, 1994.
[19] K. Roy, S. Mukhopadhyay and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proceedings of IEEE*, Vol. 91, pp. 305 – 327, 2003.
[20] J. A. Butts and G. S. Sohi, "A static power model for architects," *Proceedings of IEEE/ACM International Symposium on Microarchitecture (Micro33)*, pp. 191 – 201, 2000.
[21] E. J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM Journal of Research and Development*, Vol. 46, pp. 169 – 180, 2002.
[22] O. Semenov, A. Vassighi, M. Sachdev, A. Keshavarzi and C. F. Hawkins, "Burn-in temperature projections for sub-micron technologies," *ITC International Test Conference*, pp. 95 – 104, 2003.
[23] M. Schulz, "The end of the road for silicon?" Nature, Vol. 399, pp. 729 – 730, 1999.
[24] K. Banerjee, S. 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," *IEDM*, pp. 36.7.1 – 36.7.4, 2003.
[25] S. Borkar, "Low power design challenges for the decade" *Design Automation Challenges Conference (DAC 2001)*, pp. 293 – 296, 2001.
[26]http://www.intel.com/pressroom/kits/quickreffam.htm
[27] R. Saleh, G. Lim, T. Kodawaki, and K. Uchiyama, "Trends in Low Power Digital System-on-Chip Designs," *International Symposium on Quality Electronic Design*, pp. 373 – 378, 2002.
[28] A. Basu, S-C. Lin, V. Wason, A. Mehrotra, and K. Banerjee, "Simultaneous Optimization of Supply and Threshold Voltages for Low-Power and High-Performance Circuits in the Leakage Dominant Era," *41st Annual ACM IEEE Design Automation Conference*, pp. 884 – 887, 2004.
[29] G. Sery, S. Borkar and V. De, "Life is CMOS: Why chase the life after?" *Proceedings of DAC 2002*, pp. 78 – 83, 2002.
[30] I. Sauciuc, R. Prasher, J. Y. Chang, H. Erturk, G. Chrysler, C. P. Chiu and R. Mahajan , "Thermal performance and key challenges for future CPU cooling technologies," *Proceedings of the ASME/Pacific Rim Technical Conference and Exhibition on Integration and Packaging of Micro, Nano, and Electronic Systems (InterPACK '05)*, IPACK2005-73242, 2005.

[31] I. Sauciuc, H. Erturk, J. Williams, G. Chrysler and R. Mahajan, "Integration of thermoelectric and phase change (liquid/vapor) devices with application to CPU cooling," IMAPS-ATW, Palo Alto, CA, 2004.
[32] R. C. Chu, R. E. Simons, M. J. Ellsworth and R. R. Schmidt, and V. Cozzolino, "Review of cooling technologies for computer products," *IEEE Transactions Device and Materials Reliability*, Vol.4, pp. 568 – 585, 2004.
[33] I. Aller, K. Bernstein, U. Ghoshal, H. Schettler, S. Schuster, Y. Taur and D. Torreiter, "CMOS Circuit technology for sub-ambient temperature operation," *IEEE International Solid-State Circuits Conference*, pp. 214 – 215, 2000.
[34] W. H. Henkels, D. Wen, R. L. Mohler, R. L. Fahler, T. J. Bucelot, C. W. Long, J. A. Bracchitta, W. J. Cote, G. B. Bronner, Y. Taur and R. H. Dennard, "A 4-MB low-temperature DRAM," *IEEE Journal of Solid-State Circuits*, Vol. 26, pp 1519 – 1529, 1991.
[35] A. Vassighi, A. Keshavarzi, S. Narendra, G. Schrom, Y. Ye, S. Lee, G. Chrysler, M. Sachdev and V. De, "Design optimizations for microprocessors at low temperature," *DAC'2004*, pp. 2-5, 2004.
[36] T. C. Harman, P. J. Taylor, M. P. Walsh and B. E. LaForge, "Quantum dot superlattice thermoelectric materials and devices," *Science*, Vo. 297, pp. 2229 – 232, 2002.
[37] N. W. Ashcroft and N. D. Mermin, Solid State Physics, Thomson Learning, 1976.
[38] M. D. Ulrich, P. A. Barnes and C. B. Vining, "Comparison of solid-state thermionic refrigeration with thermoelectric refrigeration," *Journal of Applied Physics*, Vol. 90, pp. 1625 – 1631, 2001.
[39] Y. Hishinuma, T.H. Geballe, B.Y. Moyzhes, and T.W. Kenny, "refrigeration by combined tunneling and thermionic emission in vacuum: Use of nanometer scale design," *Applied Physics Letters*, vol. 78, pp. 2572-2574, 2001.
[40] G. S. Nolas and H. J. Goldsmid, "A comparison of projected thermoelectric and thermionic refrigerators," *Journal of Applied Physics*, Vol. 85, pp. 4066 – 4071, 1999.
[41] M. D. Ulrich, P. A. Barnes and C. B. Vining, "Comparison of solid-state thermionic refrigeration with thermoelectric refrigeration," *Journal of Applied Physics*, Vol. 90, pp. 1625 – 1631.
[42] J. H. Goldsmid, Electronic Refrigeration, Pion, 1986.
[43] G. Chen, and A. Shakhouri, "Heat transfer in nanostructures for solid-state energy conversion," *Journal of Heat Transfer*, Vol. 124, pp. 242 – 252, 2002.
[44] G. Chen, M. S. Dresselhaus, J. P. Fleurial and T. Caillat, "Recent developments in thermoelectric materials," International Material Reviews, Vol. 48, pp. 1 – 22, 2003.
[45] Hodes, M., "On one-dimensional analysis of thermoelectric modules (TEMs)," *IEEE Transactions on Components and Packaging Technologies*, Vol. 28, pp. 218 – 229, 2005.
[46] L. I. Anatychuk and O. J. Luste, "Physical principles of microminiaturization in thermoelectricity," *15th International Conference on Thermoelectrics*, pp. 279 – 282, 1996.
[47] O. Yamashita, "Effect of metal electrode on Seebeck coefficient of p- and n-type silicon thermoelectrics," *Journal of Applied Physics*, Vol. 95, pp. 178 – 183, 2004.
[48] L. W. da Silva and M. Kaviany, "Micro-thermoelectric cooler: interfacial effects on thermal and electrical transport," *International Journal of Heat and Mass Transfer*, Vol. 47, p. 2417 – 2435, 2003.
[49] V. Semeniouk and J. P. Fleurial, "Novel high performance thermoelectric microcoolers with diamond substrates," *16th International Conference on Thermoelectrics*, pp. 683 – 686, 1997.
[50] J. P. Fleurial, A. Borshchevsky, M. A. Ryan, W. Phillips, E. Kolawa, T. Kacisch and R. Ewell, "Thermoelectric microcoolers for thermal management applications," *16th International Conference on Thermoelectrics*, pp. 641 – 645, 1997.
[51] U. Ghoshal, Y. S. Ju, A. Miner and M. B. Ketchen, "Advanced electronic microcoolers," *18th International Conference on Thermoelectrics*, pp. 113 – 116, 1999.
[52] G. Min and D. M. Rowe, "Improved model for calculating the coefficient of performance of a Peltier module," *Energy Conversion and Management*, Vol. 41, pp. 163 – 171, 2000.
[53] Y. S. Ju and U. Ghoshal, "Study of interface effects in thermoelectric microrefrigerators," *Journal of Applied Physics*, Vol. 88, pp. 4135 – 4139, 2000.
[54] R. E. Simons, M. J. Ellsworth, and R. C. Chu, "An assessment of module cooling enhancement with thermoelectric coolers," *Journal of Heat Transfer*, Vol. 127, pp. 76 – 84, 2005.

Table 1. Parameter values and trends [25,26,27,28].

Parameters	**Initial Value**	**Trends**
Capacitance (fF)	1.9	0.7x
Voltage (V)	1.5	0.85x
Average number of transistors (millions)	55	2x
Frequency (GHz)	2.2	1.5x
Activity factor	0.1	
Die size (cm^2)	1.4	

Table 2. Comparison of electronic refrigeration technologies [38-42].

Technology	Physical Mechanism	Advantages	Disadvantages	Remarks
Thermoelectric	Current passed through a two dissimilar metals (or) semiconductors that are connected to each other at two junctions (Peltier junction)	- Theory and values well established - Flexibility in fabrication routes (Sputtering, Vapor deposition, etc)	- Joule heating - Heat conduction	- Thermo-tunneling has the highest theoretical COP compared to thermoelectrics & solid-state thermionic - Thermoelectric & solid-state thermionic $\square\beta_{TI}/\beta=(d/\square\lambda\pi^{0.5})$ (d – barrier width; λ - mean-free path); For $d < \lambda$, thermoelectric coolers are superior than solid-state thermionic devices
Solid-state Thermionic	Electrons passing through a barrier layer from the emitter to the collector cause cooling at the emitter-barrier junction (ballistic transport of electrons)	- High ZT for thin films - In ideal device, Joule heating is absent	- Very complex material system and may not be easy to controllably fabricate - Often slow fabrication	
Thermionic (Thermotunneling)	Selective evaporation emission of hot electrons over a barrier layer combined with quantum mechanical electron tunneling (vacuum gap) from cathode to anode	- Theoretical efficiency (COP/COP_{Carnot}) ~ 0.7 - No ohmic resistance - Radiation is the only mode of heat transfer	- Not easy to implement for the case of chip cooling due to small gaps and requirement of vacuum - Material work function values limits practical devices (>500 K)	

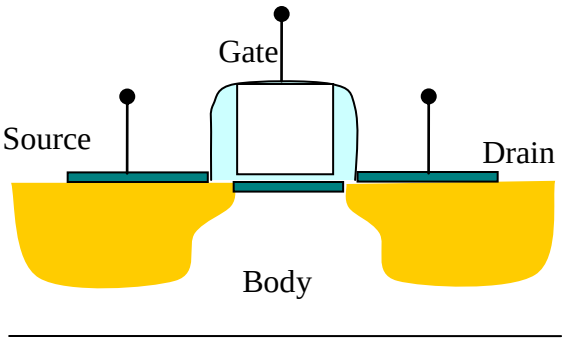


Figure 1. Schematic diagram of MOS transistor.

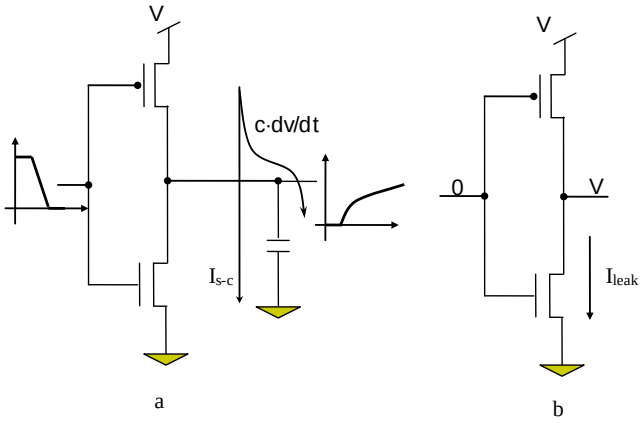


Figure 2. Schematic diagram of (a) dynamic and (b) static power dissipation mechanisms in CMOS technologies. The dynamic power dissipation is due to the charging of capacitance. It is noted that the input is zero for static power dissipation.

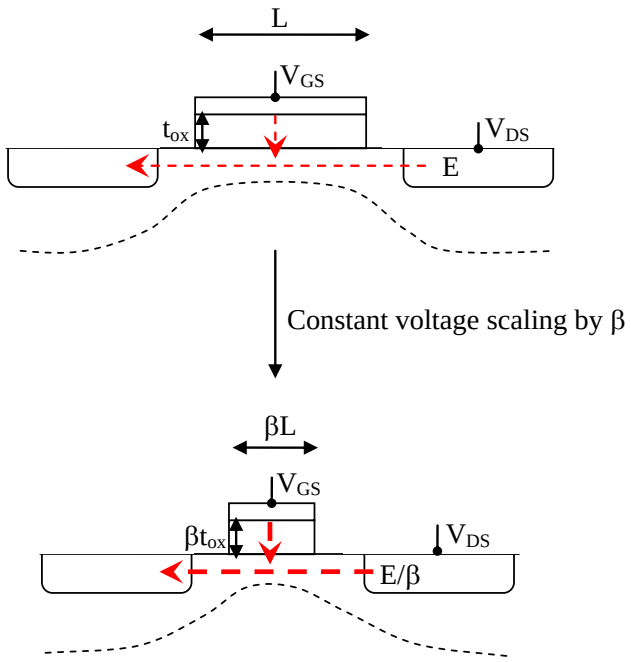


Figure 3. Schematic illustration of constant voltage scaling. The electric field increases as the technology is scaled, and the channel length and gate oxide thickness decrease while voltage remains constant.

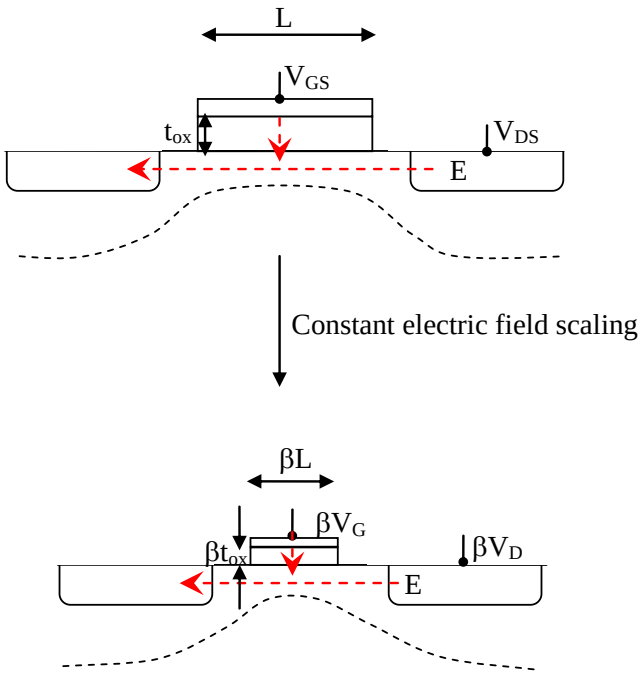


Figure 4. Schematic illustration of constant electric field scaling. The electric field remains constant while the linear dimensions and voltage decrease with each generation.

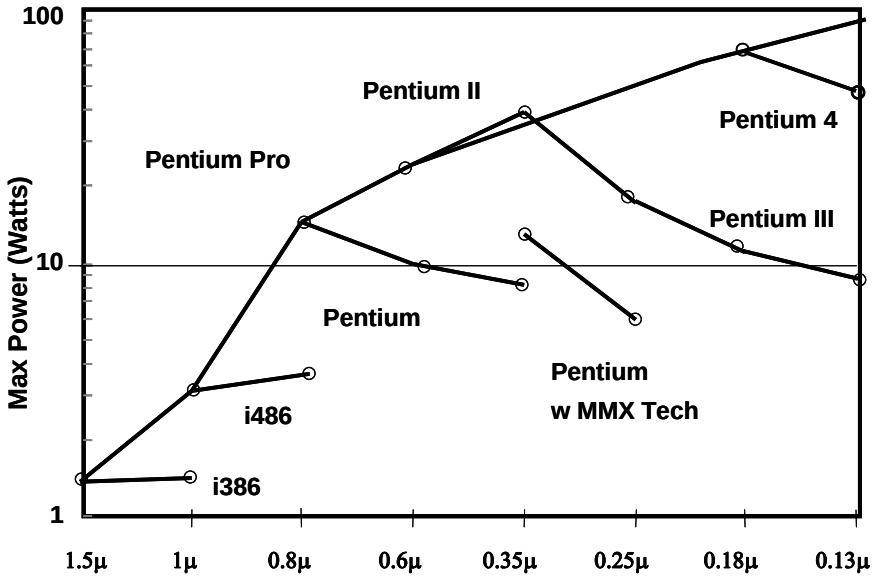
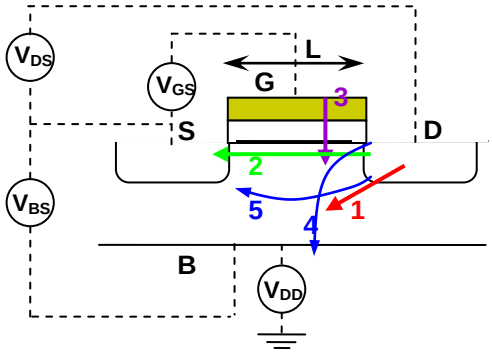


Figure 5. Maximum power dissipation of several microprocessors [17].



1 – pn Junction leakage current
2 – Sub-threshold leakage current
3 – Gate-oxide tunneling
4 – Gate induced drain leakage current
5 – Punchthrough

Figure 6. Schematic illustration of the various leakage currents in MOS transistors. S, G, B, D and L in the figure are source, gate, body, drain terminals and channel length, respectively.

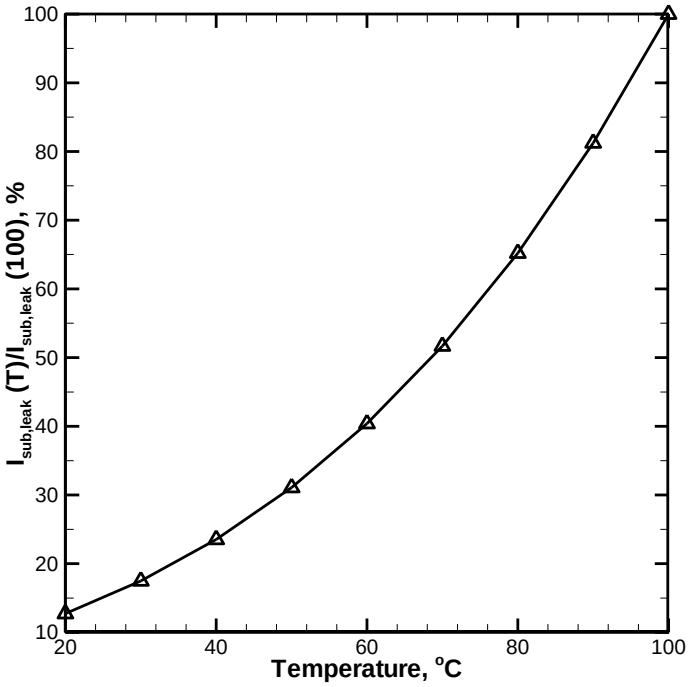


Figure 7. Sub-threshold leakage power normalized with respect to temperature as a function of temperature.

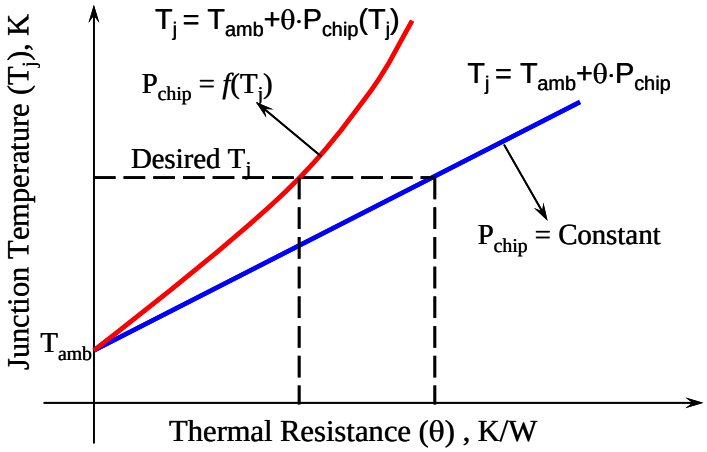
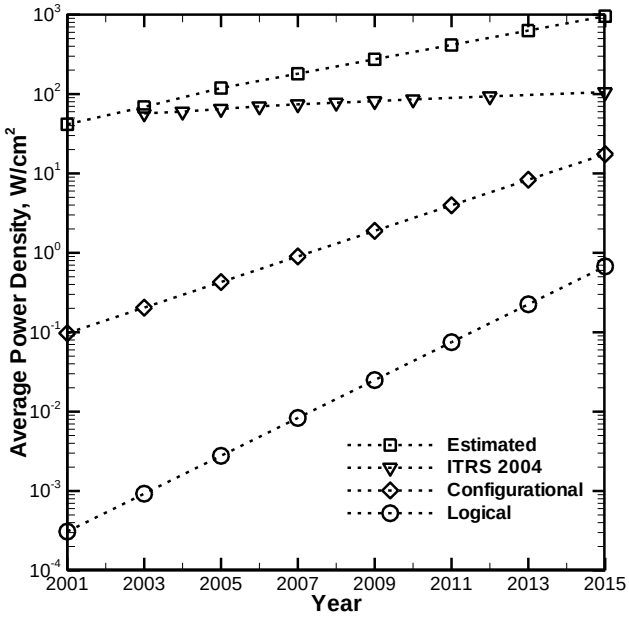
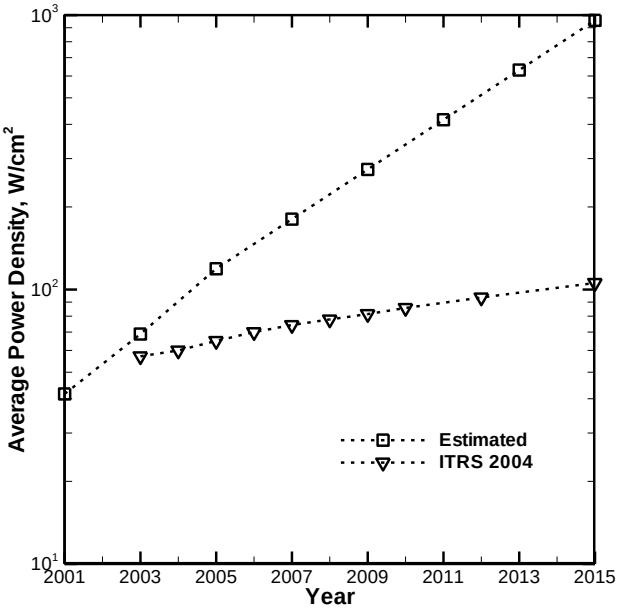


Figure 8. Schematic diagram illustrating the electro-thermal coupling of junction temperature and total thermal resistance [24].



(a)



(b)

Figure 9. Estimated evolution of average power density (a) with and (b) without the lower limits on the power dissipation (projected power density is based on Equation (4)). See section 2.5.

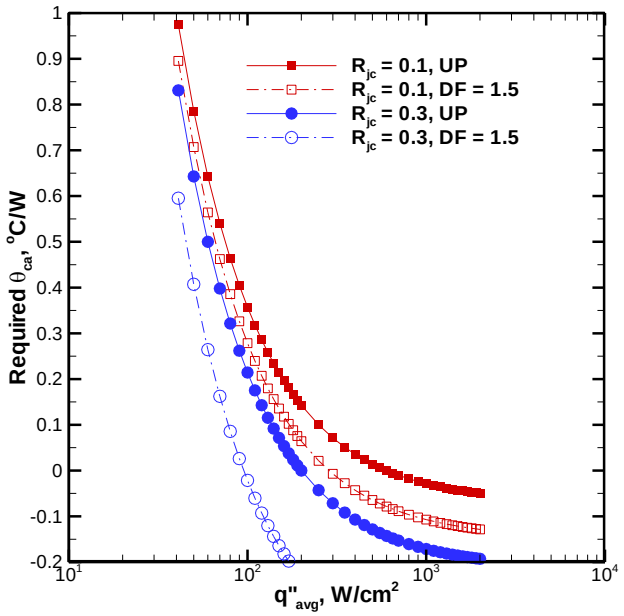


Figure 10. Evolution of external heat sink requirements based on the projected average power density trends (UP – Uniform Power and DF – Density Factor). The ambient temperature is assumed to be 40°C in this calculation.

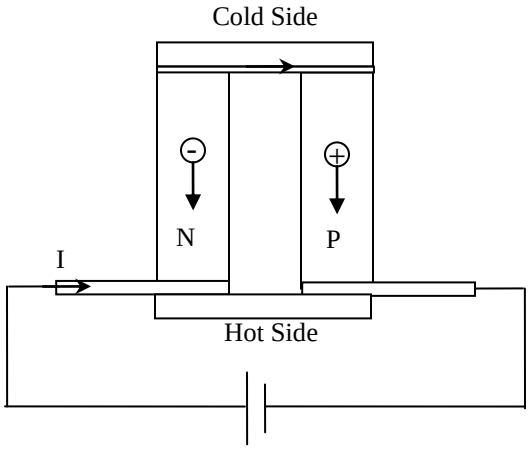


Figure 11. Schematic diagram of a solid-state thermoelectric cooler.

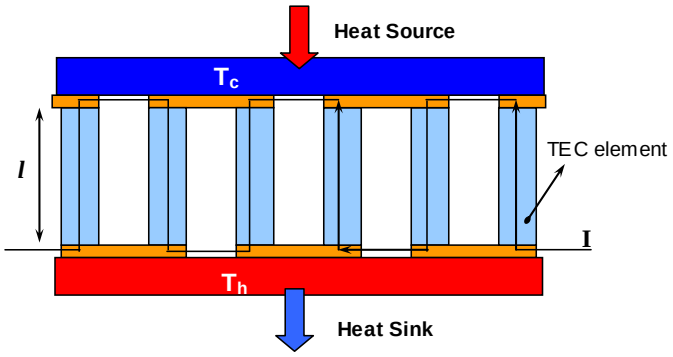


Figure 12. Schematic diagram of an ideal single-stage thermoelectric cooling module.

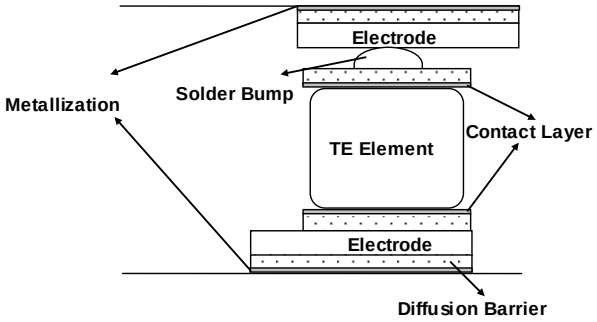


Figure 13. Schematic diagram of a thermoelectric device with various sources of electrical contact resistance (adapted from [50]).

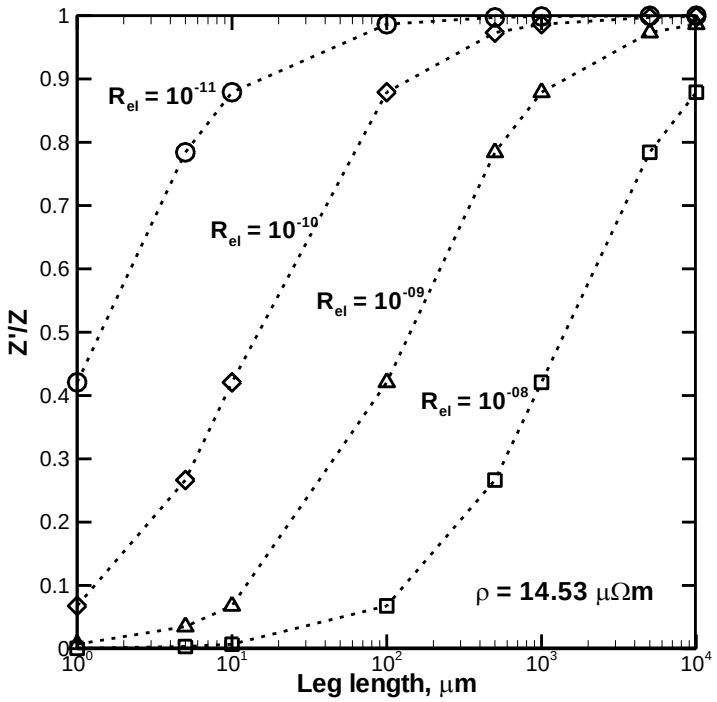


Figure 14. Effect of electrical contact resistance (R_{el}) on operating figure of merit of the TEC module.

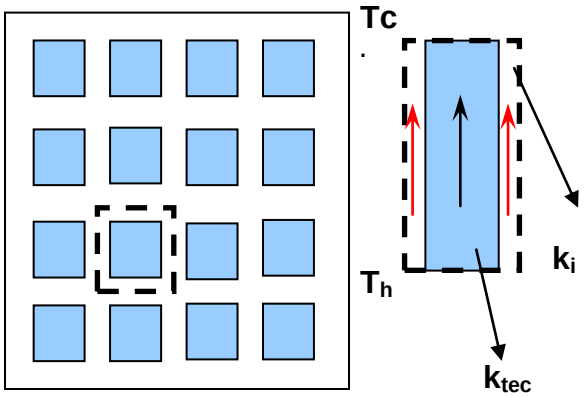


Figure 15. Schematic of back flow of heat through the interstitial and thermoelectric element along with the unit-cell model considered.

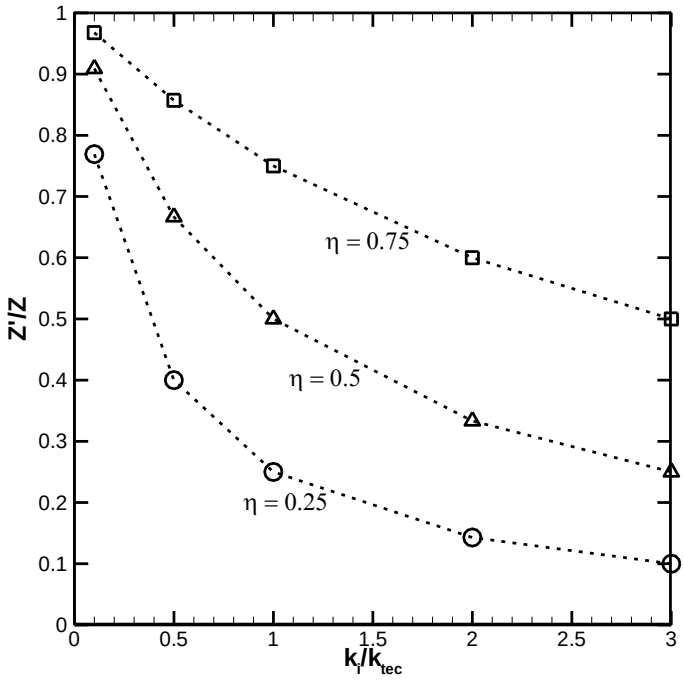


Figure 16. Effect of back conduction through the interstitial material on the operating figure of merit (Z') of the TEC module.