Analysis and Implications of IC Cooling for Deep Nanometer Scale CMOS Technologies

Sheng-Chih Lin, Ravi Mahajan[1], Vivek De[2] and Kaustav Banerjee
Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106
[1]Assembly Technology and Development, Intel Corp., Chandler, AZ, 85226 [2]Circuit Research Labs, Intel Corp., Hillsboro, OR, 97124

Abstract

Cooled chip operation is being seriously evaluated as a practical technique for boosting the performance of high-end microprocessors. This paper presents, for the first time, a comprehensive analysis of chip cooling for various nanometer scale bulk-CMOS and Silicon-On-Insulator (SOI) technologies. Unlike all previous work, our 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 cooling always gives performance gain at the device or circuit level, it is shown that system level power defines a temperature limit beyond which cooling gives diminishing returns and the associated cost 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.

I. Introduction

Power dissipation and related thermal problems are beginning to severely impact the scalability of high-performance IC products including microprocessors (**Fig. 1**) [1]. These problems are exacerbated by the fact that for deep nanometer scale CMOS technologies, leakage power forms a significant fraction of total chip power and hot-spot induced substrate temperature gradients exist in high-performance ICs [2]. The increase in subthreshold leakage power arises due to the fact that power supply (V_{dd}) scaling necessitates threshold voltage (V_{th}) scaling (**Fig. 2**). Additionally, parameter variations [2] are leading to significant increase in leakage power as shown in **Fig. 3**. Prior work in chip cooling has primarily focused on sub-ambient [3-4] or cryogenic [5-6] temperatures. However, the practicality of such operating temperatures is questionable. Moreover, these analyses were carried out at the device or circuit level 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 work presents a comprehensive evaluation of different cooling strategies applied to scaled technologies in the range of realistic operating temperatures for microprocessors. In addition to performance, the benefits of cooling are also quantified from a system level power dissipation point of view.

II. Analysis of cooling at the device and circuit 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 because of the reduction of carrier scattering caused by thermal vibrations of the semiconductor crystal lattice. Thin-body SOI CMOS is an attractive alternative to bulk CMOS due to superior electrostatics and speed (lower junction capacitance), as well as better subthreshold-swing, and latch-up/SER immunity. **Fig. 4** compares the I-V characteristics of bulk and floating-body partially-depleted SOI type transistors (PD-SOI) at different temperatures. **Fig. 5** shows that the drive (drain) current capability increases with transistor scaling. Due to higher carrier mobility at lower temperature, it can be observed from **Fig. 4** and **5** that higher drive current can be achieved by cooled operation across all technology nodes. In **Fig. 6**, while it is evident that drive current increases at lower temperature 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_{BS}) of PD-SOI increases as temperature decreases [7], which causes a smaller increase in the saturated threshold voltage of PD-SOI type transistors compared to bulk transistors at lower temperature, as shown in **Fig. 7**. Thus, the enhancement of drive current of PD-SOI transistors at low temperature is higher than that in bulk transistors.

Subthreshold leakage, the main leakage contributor, is highly temperature dependent. **Fig. 8** shows the device subthreshold characteristics for both bulk and SOI transistors. Clearly the device under cooled operation results in a steeper subthreshold slope than under normal operation. It can be observed from **Fig. 9** that subthreshold leakage *increases* exponentially with transistor scaling while **Fig. 10** shows that the subthreshold 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**, lowering temperature significantly improves the I_{on} to I_{off} ratio for bulk as well as for SOI at all technology nodes. **Fig. 12** shows the amount of cooling that will be needed in order to meet ITRS prescribed requirement for saturation drive current without redesigning [8].

Fig. 13 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*). Although the threshold voltage increases at lower operating temperature and partially offsets the performance improvement gained from the higher carrier mobility, cooling still gives net improvement of performance. Redesign strategies can be used to further enhance the benefit of cooling. For example, adjusting the threshold voltage by body-biasing (e.g., applying forward body bias to lower threshold voltage) to maintain the same threshold voltage (*same* V_{th}) at lower temperatures shows a higher net performance improvement mainly due to the higher mobility. Further lowering of the threshold voltage to maintain the same off-state leakage current (*same leakage*) results in maximum improvement in performance that can be achieved by cooling. These redesign strategies become increasingly effective in exploiting the benefits from cooling as technology scales.

Higher device drive (drain) current capability at lower temperature enhances the circuit performance. The distribution of gate propagation delay of a 9-stage inverter chain (30 samples) under different operating temperature was estimated by Monte Carlo analysis. As shown in **Fig. 14**, the mean value of gate propagation delay improves by 9% at the lower temperature. Moreover, the variation in the circuit performance due to channel length variation can also be mitigated (variance reduces by 13%). 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. Reliability of interconnects (electromigration) and inter-layer dielectrics (TDDB) also improves due to cooling. For semi-global and global wires, more aggressive interconnect scaling (narrower wire and/or smaller aspect ratio) is allowed to improve interconnect RC delay without degrading EM reliability and inductive effects [9]. Interconnect

0-7803-9269-8/05/$20.00 (c) 2005 IEEE

with smaller aspect ratio will also improve the routability of the chip and enhance the rate at which bits can be transmitted per unit chip edge, i.e., bandwidth [10]. Also, the number and size of repeaters needed along global interconnects can be reduced leading to lower power dissipation [11].

Fig. 15 shows the increase of active power dissipation (due to higher frequency of operation) at different operating temperatures for the same design scenarios in **Fig. 13**. Although lower temperature enhances the performance and the 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 active (P_{active}) and leakage ($P_{leakage}$) power dissipation at the device level and ignoring the additional cooling power ($P_{cooling}$) required to achieve lower junction temperature. The aspect of system level power dissipation is addressed in the following section.

III. Analysis of cooling at the system level

A variety of cooling technologies has been proposed for improving performance of future high-performance ICs [12]. It has been shown that system performance can not be correctly evaluated without considering electrothermal couplings, junction temperature, and associated cooling power [13]. The inset of **Fig. 16(a)** shows the improvement of electrothermally-aware system level performance while applying cooling. As expected, similar to the analyses at 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 if electrothermal couplings are neglected and if they do not comprehend the dynamic nature of system performance improvement with power.

Fig. 16(a) shows the leakage power dissipation for two identical test microprocessor designs at different technology nodes under the application of active cooling, using the methodology described in [13]. It can be observed that leakage power dissipation decreases significantly as more cooling power is applied and the reduction of leakage power becomes greater as technology scales. **Fig. 16(b, c)** show that the chip power ($P_{chip}=P_{active}+P_{leakage}$) decreases as more cooling is applied mainly as a result of decreasing leakage power as shown in **Fig. 16(a)**. The total system power ($P_{system}=P_{chip}+P_{cooling}$), however, decreases only as long as the savings in chip power dissipation remains greater than the additional power required for cooling. Hence, it can be observed from **Fig. 16(b, c)** that there is a clear minimum point in the curve of total system power that determines the practical limit beyond which further cooling does not lead to any overall power savings and the limit occurs at an increasingly lower temperature as technology scales. It can thus be concluded that, as technology scales, the benefit that can be derived from cooling increases. **Fig. 17** demonstrates that the practical limit of cooling can be further extended towards a lower operating temperature (and hence higher performance) by enhancing the cooling efficiency.

Design parameters such as supply voltage can also affect the usefulness of cooling in improving chip performance and power dissipation. **Fig. 18** compares the total system power dissipation 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. At lower supply voltages, although active cooling results in lower operating temperature, the total system power dissipation is higher than that with passive cooling, as expected. However, at higher supply voltage (V_{dd} = 1.2V) active cooling not only leads to lower operating temperature but also results in lower total system power consumption as compared to passive cooling. This is because the extra power spent on cooling is lower than the corresponding saving in leakage power.

IV. Effective cooling strategy for hot-spot management

In order to comprehend the impact of cooling on thermal gradients and hot-spots, a self-consistent electrothermal substrate thermal profile generating methodology has been developed as shown in **Fig. 19**. Thermal parameters of a typical microprocessor package structure, Flip-Chip Land-Grid-Array (FC-LGA) and a socket that interfaces with the printed-circuit board (PCB) were used for the thermal profile estimation. This methodology is based on the solutions of the parabolic heat partial differential equations (both vertical and lateral heat transfer are considered) incorporating an electrothermally-aware self-consistent approach [13]. The parabolic PDEs were solved using the Alternating-Direction-Implicit (ADI) method [14] for achieving high computation efficiency.

An example chip design (die size: 10 mm × 10 mm) with power densities per functional block is shown in **Fig. 20(a)**. The spatial substrate temperature profile, **Fig. 20(b)**, shows several hot-spots and the highest junction temperature is around 133°C. Although the results shown here are specific to the above mentioned IC, the conclusions drawn are more generic. **Fig. 21** shows the effect of applying global and localized cooling strategies on hot-spot management. As shown in **Fig. 21(a)**, a lower junction-to-ambient thermal resistance (θ_{ja}) reduces the maximum junction temperature by applying global cooling (through better interface material, higher cooling efficiency, etc.). However, on-chip hot-spots and thermal gradients still remain. On the other hand, localized cooling solutions such as local spray cooling, thin-film thermoelectric coolers, can be applied to electronic application to eliminate the hot-spots. For example, if two thin-film thermoelectric coolers can be placed on the backside of the wafer below the locations of hot-spots, as shown in **Fig. 21(b)**, it can effectively eliminate the targeted hot-spots.

V. 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 partially-depleted SOI based technologies are more responsive to cooling. It is also demonstrated that while cooling always gives performance gain 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.

Acknowledgement

This work was supported by Intel Corporation and the University of California-MICRO program.

References

[1] P. Gelsinger, *DAC* Keynote 2004.
[2] S. Borkar, et al., *DAC*, pp. 338 –342, 2003.
[3] B. Yu et al., *VLSI-TSA*, pp. 23-25, 2001.
[4] I. Aller et al., *ISSCC*, pp. 214-215, 2000.
[5] J. Sun et al., *TED*, pp. 19-27, 1987.
[6] F. Gaensslen et al., *TED*, pp. 218-306, 1977.
[7] M.M. Pelella et al., *International SOI conference*, pp. 147-148, 1998.
[8] International Technology Roadmap for Semiconductors (ITRS) 2004.
[9] K. Banerjee and A. Mehrotra, *TCAD*, pp. 904-915, 2002.
[10] M. L. Mui et al., *TED*, pp. 195-203, 2004.
[11] K. Banerjee and A. Mehrotra, *TED*, pp. 2001-2007, 2002.
[12] R. Viswanath et al., *Intel Technology Journal*, Q3, 2000.
[13] K. Banerjee et al., *IEDM*, pp. 887-890, 2003.
[14] J. Douglas et al, *Trans. American Mathematical Society*, pp. 421-439, 1956.

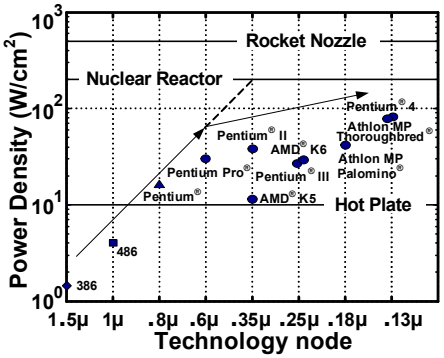


Fig. 1: Power density trends for high performance microprocessors as technology scales.

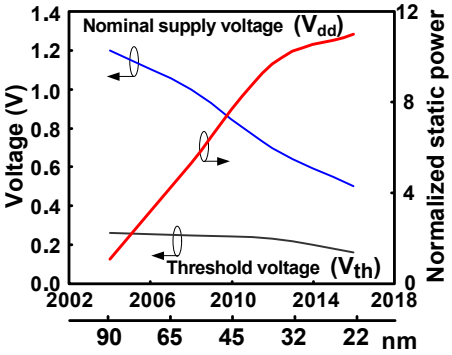


Fig. 2: Nominal supply voltage, threshold voltage and static power based on ITRS'04, as technology scales.

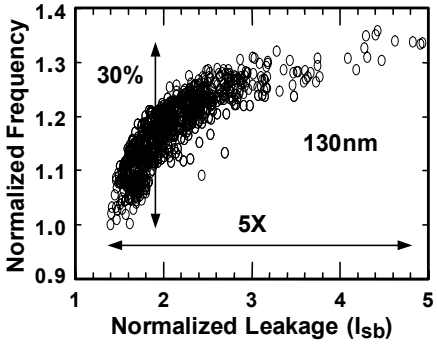


Fig. 3: Distributions of frequency and standby leakage current for different microprocessors on a single wafer.

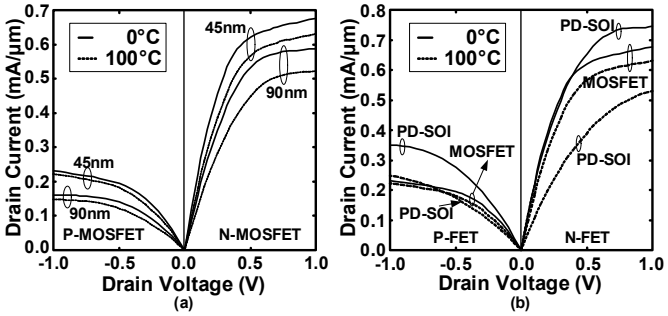


Fig. 4: Device I-V characteristics at different temperatures. **(a)** Bulk MOSFET for 45 nm and 90 nm effective channel length. **(b)** Bulk MOSFET with effective channel length 45 nm vs. Partially-Depleted SOI with effective channel length 120 nm.

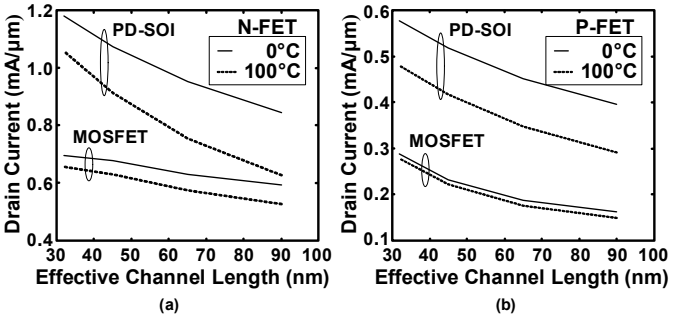


Fig. 5: Drive current for (a) N-FET and (b) P-FET as a function of effective channel length at 100°C and 0°C (cooled operation).

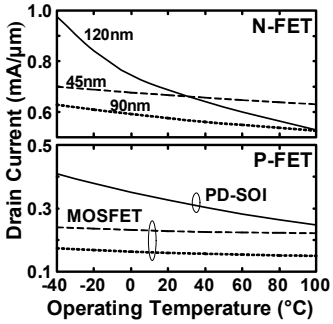


Fig. 6: Drive (drain) current for N-FET (top) and P-FET (bottom) as a function of operating temperature, at different technology nodes.

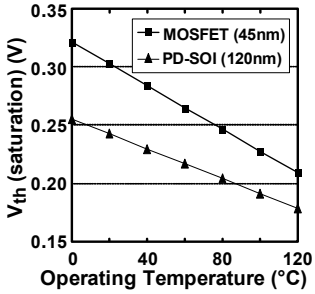


Fig. 7: The increasing rate of saturated threshold voltage for bulk MOSFET (0.9mV/°C) is larger than PD-SOI type transistor (0.6mV/°C).

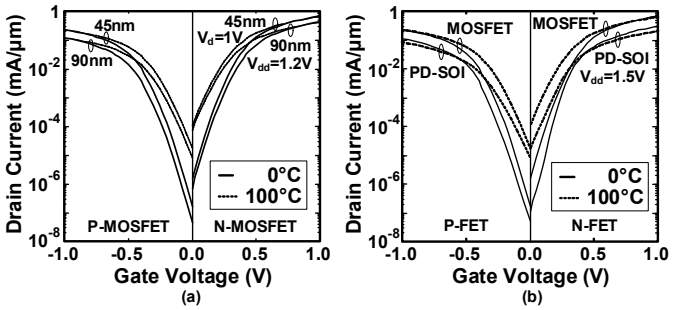


Fig. 8: Device subthreshold characteristics at different temperatures. **(a)** Bulk MOSFET for 45 nm and 90 nm effective channel length. **(b)** Bulk MOSFET with effective channel length 45 nm vs. Partially-Depleted SOI with effective channel length 120 nm.

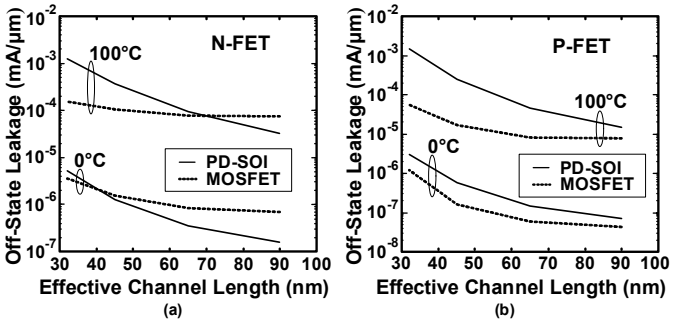


Fig. 9: Off-state leakage current for **(a)** N-FET and **(b)** P-FET with shrinking device channel length. Off-state leakage current decreases significantly under cooled operation.

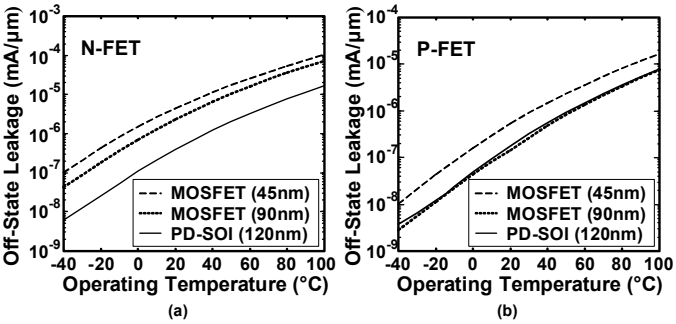


Fig. 10: Off-state leakage current for **(a)** N-FET and **(b)** P-FET as a function of temperature for different technology nodes.

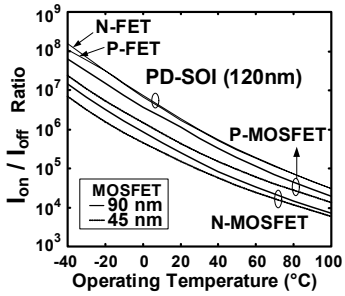


Fig. 11: Ratio of drive current to leakage current as a function of temperature at different technology nodes.

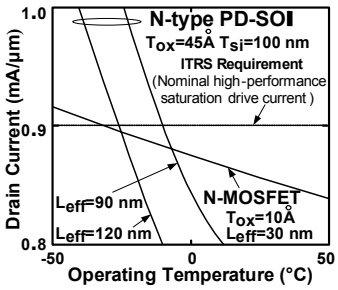


Fig. 12: ITRS requirement for N-FET saturation drive current can be achieved by lowering operation temperature around -30°C without redesign.

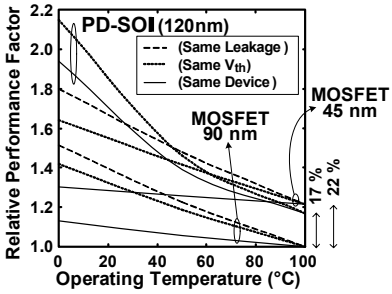


Fig. 13: Relative performance factor for three scenarios at different operating temperatures, normalized to the values at 100°C.

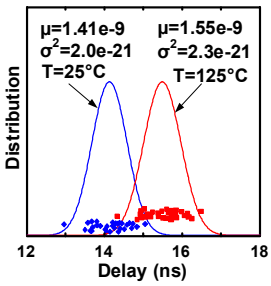


Fig. 14: Monte Carlo analysis of the propagation delay of a 9-stage inverter chain. The distribution of the delay is indicated by "♦" and "■" for 25°C and 125°C respectively.

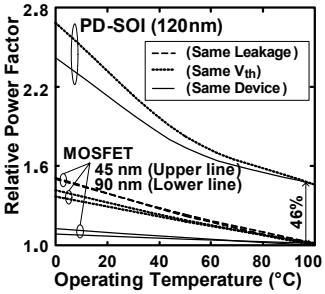


Fig. 15: Relative power factor for three scenarios at different operating temperatures, normalized to the values at 100°C.

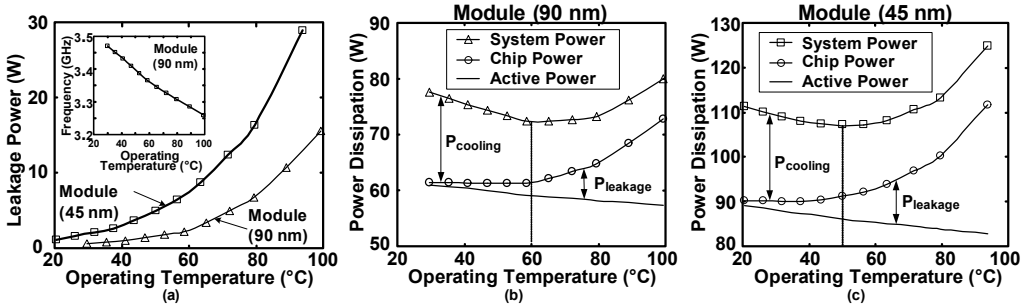


Fig. 16: Electrothermally-aware system level evaluation of power dissipation. A minimum P_{system} determines the practical limit beyond which further cooling does not lead to any power saving. **(a)** $P_{leakage}$ as a function of operating temperature. The inset shows chip frequency increases as operating temperature decreases. **(b)** Minimum P_{system} (90 nm) is around 60°C. **(c)** Minimum P_{system} (45 nm) is around 50°C.

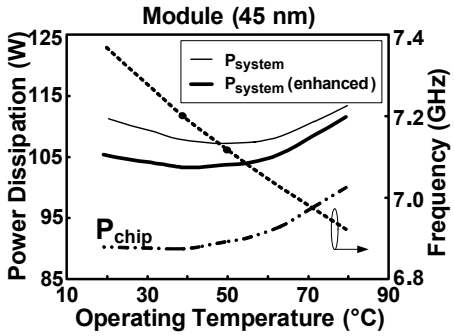


Fig. 17: Electrothermally-aware system level evaluation of power dissipation with 70% enhanced cooling efficiency. The minimum point of P_{system} moves towards a lower temperature which leads to a higher performance.

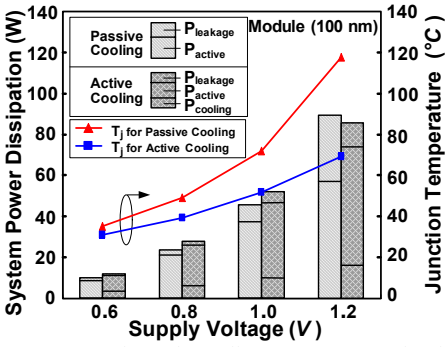


Fig. 18: Electrothermally-aware system level evaluation of power dissipation and junction temperature for passive and active cooling. For the case V_{dd} = 1.2V, P_{system} for active cooling is less than that for passive cooling.

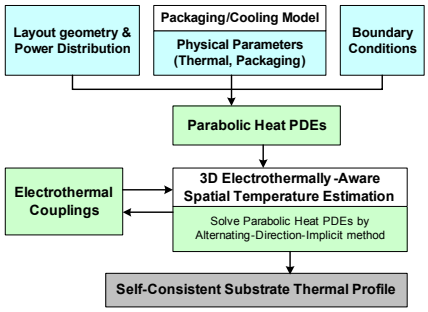
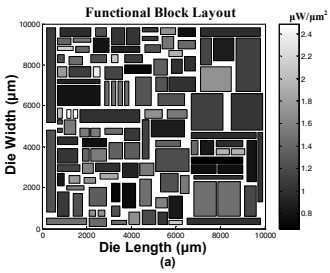


Fig. 19: Overview of the electrothermally-aware substrate thermal profile generator.



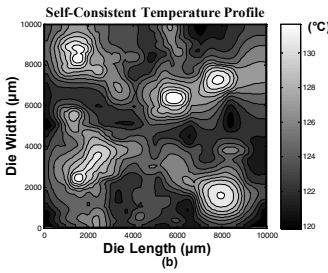
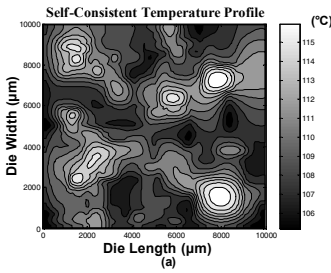


Fig. 20: (a) Functional block layout of a test chip showing power density associated with each block. Nominal total power consumption is 75 W. **(b)** Spatial substrate temperature profile of the test chip generated using the methodology described in **Fig. 19** (junction-to-ambient thermal resistance θ_{ja}=1.1 °C/W). Five hot-spots can be observed. The highest temperature (T_{max}) is around 133°C.



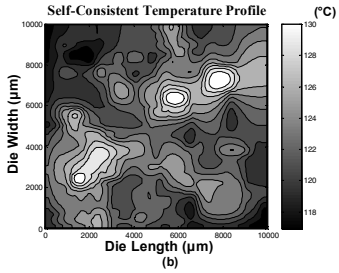


Fig. 21: (a) Spatial substrate temperature profile of the test chip generated using the methodology described in **Fig. 19**. θ_{ja} reduces 20% by applying global cooling. Although T_{max} (117°C) decreases, the hot-spots remain. **(b)** Temperature profile of the test chip after integrating two thin-film thermoelectric coolers at the top-left and bottom-right hot-spots. Only three hotspots can be observed.