High-Performance Heat Sinking for VLSI

D. B. TUCKERMAN AND R. F. W. PEASE

Abstract—**The problem of achieving compact, high-performance forced liquid cooling of planar integrated circuits has been investigated. The convective heat-transfer coefficient h between the substrate and the coolant was found to be the primary impediment to achieving low thermal resistance. For laminar flow in confined channels, h scales inversely with channel width, making microscopic channels desirable. The coolant viscosity determines the minimum practical channel width. The use of high-aspect ratio channels to increase surface area will, to an extent, further reduce thermal resistance. Based on these considerations, a new, very compact, water-cooled integral heat sink for silicon integrated circuits has been designed and tested. At a power density of 790 W/cm², a maximum substrate temperature rise of 71°C above the input water temperature was measured, in good agreement with theory. By allowing such high power densities, the heat sink may greatly enhance the feasibility of ultrahigh-speed VLSI circuits.**

Introduction

THE advent of systems employing high-speed, high-density, very-large-scale integrated (VLSI) circuits implies the requirement for effective and compact heat removal. For example, high-speed digital circuits employing submicron channel lengths yet dissipating 1 mW per gate have been reported recently [1]. A VLSI circuit containing 10^5 such gates would thus dissipate 100 W. Conventional IC packages typically have thermal resistances of 50°C/W and hence would be totally unsuitable for such circuits.

Although an isolated chip dissipating 100 W could be cooled by forced-air convection, an array of such chips (closely spaced to minimize propagation delays) presents a far more difficult cooling problem because of the large size ($\gtrsim 10$ cm on a side) of high-performance forced-air heat exchangers. Liquid cooling promises to be a more compact arrangement and its use has been reported recently for cooling the central processing unit of a large computing system [2]. In that system, heat is conducted through aluminum pistons spring-loaded onto the back of each chip, through cylinder walls surrounding the pistons, and into a heat exchanger. The thermal resistance of such a package allows a dissipation of about 3 or 4 W per chip. Even more compact configurations have been proposed by integrating the heat exchanger with the silicon chip [3].

There have been suggestions that physical limits of heat-transfer technology will limit the power density of arrays of planar circuits to 20 W/cm² or so [4]. In this letter we show that by scaling liquid-cooled heat-exchanger technology to microscopic dimensions, circuit power densities of more than 1000 W/cm² should be feasible. To demonstrate these principles, we have constructed a very compact water-cooled heat sink which is an integral part of the silicon substrate, which exhibits a maximum thermal resistance of 0.09°C/W over 1 cm² area, and which has been tested up to 790 W/cm².

Manuscript received March 10, 1981; revised March 31, 1981.
The authors are with Stanford Electronics Laboratories, Stanford, CA 94305.

Theory

The performance of a heat sink is measured by its thermal resistance $\theta = \Delta T/\dot{Q}$, where ΔT is the temperature rise of the circuit above the input coolant temperature (often room temperature) and $\dot{Q}$ is the dissipated power. In forced-convection cooling, θ is nearly independent of power level. Because semiconductor ICs typically have maximum operating temperatures of $\Delta T_{\max} = 50$°C to 100°C above room temperature, thermal resistance determines the maximum power at which an IC can operate. In general θ is the sum of three components: θ_{cond}, due to conduction from the circuits through the substrate, package, and heat-sink interface; θ_{conv}, due to convection from the heat sink to the coolant fluid, and θ_{heat}, due to heating of the fluid as it absorbs energy passing through the heat exchanger.

We can make θ_{cond} very small by locating the heat exchanger (containing the flowing coolant) very near to the heat source. Fortunately silicon, the substrate used for most planar ICs, has a high thermal conductivity ($k_{\text{Si}} = 1.48$ W/°C-cm at 27°C for lightly-doped Si; about 1/3 of copper's thermal conductivity) [5]. If an IC substrate is thinned to 100 μm and its back side is in intimate thermal contact with the heat exchanger, then θ_{cond} is only 0.007°C/W for a 1-cm² circuit.

We can reduce θ_{heat} by using a coolant of high volumetric heat capacity ρC_p at a sufficiently high flow rate f ($\theta_{\text{heat}} = 1/\rho C_p f$). Water is a particularly good choice ($\rho C_p = 4.18$ J/°C-cm³), with a modest flow rate of 10 cm³/s contributing only 0.024°C/W to the thermal resistance.

Because θ_{cond} and θ_{heat} can be made very small by rather obvious means, we expect that convective thermal resistance, θ_{conv} will be the dominant consideration in high-performance heat sink design. In fact a naive approach to liquid cooling in which water simply flows over the back of a circuit substrate can result in θ_{conv} being orders of magnitude above the other thermal resistances. It is therefore necessary to examine some aspects of convective heat-transfer theory [6].

Consider a collection of n parallel channels each of length L, imbedded in a substrate of the same length L and width W. A coolant flows in each channel, absorbing a constant heat flow per unit length $\dot{Q}/nL$ from it walls (the substrate). For example, these channels might be etched directly in the back of a silicon IC chip. The use of many separate channels, rather than a single coolant flow over the entire back substrate surface, allows us to multiply the substrate surface area by a factor α. Specifically, we define α =(total surface area of channel walls in contact with

0193-8576/81/0500-0126$00.75 © 1981 IEEE

Authorized licensed use limited to: Stanford University. Downloaded on May 8, 2009 at 19:52 from IEEE Xplore. Restrictions apply.

TABLE I

Experimental values of maximum thermal resistance Θ_{max} for three integral water-cooled silicon heat sinks of specified channel with w_c and depth z, wall thickness w_w, water pressure P, and flow rate f. The heated area was approximately (1 cm) × (1 cm), and the heat sinks were tested up to a specified maximum power density $\dot{Q}$.

Expt	$w_c(\mu m)$	$w_w(\mu m)$	$z(\mu m)$	P(psi)	$f(cm^3/s)$	Θ_{max}(°C/W)	$\dot{Q}(W/cm^2)$
1	56	44	320	15	4.7	0.110	181
2	55	45	287	17	6.5	0.113	277
3	50	50	302	31	8.6	0.090	790

fluid) ÷ (area of circuit). At each cross-section along the length of the channel, we initially assume that the walls are infinitely thermally conductive so that the temperature is uniform around the perimeter. The convective heat-transfer coefficient h is then defined as $h=\dot{Q}/nLp(T_w-T_f)$, where T_w is the wall temperature, T_f is the mean fluid temperature, and p is the cross-sectional perimeter. Then $\theta_{conv}=1/hnLp=1/h\alpha LW$, so that for a given circuit area LW, we clearly want to make both h and α large. Whereas it is well known that the use of extended-surface (large-α) structures such as fins will enhance heat transfer, the importance of making h large has received less attention.

It is customary to calculate h using dimensionless groups:

$\mathrm{Nu}=hD/k_f$, the Nusselt number, a dimensionless heat-transfer coefficient;
$\mathrm{Pr}=\mu C_p/k_f$, the Prandtl number, a property of the fluid ($\mathrm{Pr}=6.4$ for water at 23°C);
$\mathrm{Re}=vD\rho/\mu$, the Reynolds number.

Here D is a "characteristic width" of the channel, defined as $D=4\cdot$(cross-sectional area)÷(perimeter p). For high-aspect ratio rectangular channels, D is equal to twice the channel width. The terms μ, k_f, ρ, C_p, and v denote respectively the viscosity, thermal conductivity, density, specific heat, and mean velocity of the coolant fluid. Noting that the channel width D is likely to be small because the channels must be very close to the circuits to minimize θ_{cond} we tentatively assume laminar flow (a valid assumption when $\mathrm{Re}\lesssim 2100$). For calculating Nu, we further assume that the flow is "fully-developed," i.e., invariant along the channel length (a good assumption if $\mathrm{Pr}\gtrsim 5$, as is the case for most liquids). Then Nu is a monotonically decreasing function of $x/(D\cdot\mathrm{Re}\cdot\mathrm{Pr})$, where x is the distance from the entrance of the channel ($0\lesssim x\lesssim L$). Asymptotic formulas are:

$$\mathrm{Nu}\propto\left(\frac{x}{D\cdot\mathrm{Re}\cdot\mathrm{Pr}}\right)^{-1/3}, \quad \text{for } x/(D\cdot\mathrm{Re}\cdot\mathrm{Pr})\ll 0.02;$$

$\mathrm{Nu}\simeq\mathrm{Nu}_\infty$, a constant, for $x/D\cdot\mathrm{Re}\cdot\mathrm{Pr})\gtrsim 0.02$ ("fully-developed temperature profile").

Not knowing *a priori* which region we are in, we conservatively assume that Nu has the minimum, asymptotic (large x) value Nu_∞; in any case the dependence of Nu on x is weak. The exact value of Nu_∞ depends on the shape of the channel cross section but is usually between 3 and 9.

Thus we approximate $h=k_f\mathrm{Nu}_\infty/D$, where Nu_∞ is between 3 and 9. This result is consistent with an intuitive model for convection in which the heat is conducted through the fluid to the middle of the channel, where it is transported away by the flow. For a given coolant fluid, clearly the only way to significantly increase h is to reduce D. Achieving very high values of h therefore requires channels of microscopic width.

The only important lower limit on channel size is set by the coolant viscosity. For a given pump pressure, the volumetric flow rate decreases rapidly as D is reduced, resulting in an increase in θ_{heat}. By assuming a practical limit on the available pressure, we can calculate an optimum channel size D which minimizes the sum of θ_{conv} and θ_{heat}. A more fundamental limit on channel size occurs when the pumping power becomes comparable to the circuit power dissipation (and hence viscous heating becomes significant), but this only occurs at impractically high pressures.

Increasing the channel aspect ratio (i.e., increasing α) can further reduce θ_{conv}. However, we had assumed infinitely-conductive channel walls; for a substrate with finite thermal conductivity, there is little benefit in increasing α beyond the point at which thermal resistance due to conduction along the length of the walls becomes comparable to convective thermal resistance.

Design

Figure 1 is a diagram of a high-performance IC heat sink which embodies the principles just discussed. The front surface of the substrate (length L, width W) contains a planar heat source (the circuits), and the back surface contains deep rectangular channels of width w_c and depth z which carry the coolant, separated by walls of width w_w. Neglecting the heat transferred at the top and bottom of the channels, the surface-area multiplication factor due to the channels is $\alpha=2z/(w_c+w_w)$. A cover plate is bonded to the back of the substrate to confine the coolant to the channels. We will neglect θ_{cond} in our discussion, because it can be made very small independently of θ_{conv} and θ_{heat} by making the substrate only slightly thicker than the channel depth z.

Recall that $\theta_{conv}=1/h\alpha LW=D/k_f\mathrm{Nu}_\infty\alpha LW$ for infinitely conductive walls. To account for a finite wall conductivity k_w (which implies a nonuniform temperature up the walls), we can multiply by a correction factor η^{-1}, where η is known as the "fin efficiency." Approximating D as $2w_c$ for high-aspect ratio

Authorized licensed use limited to: Stanford University. Downloaded on May 8, 2009 at 19:52 from IEEE Xplore. Restrictions apply.

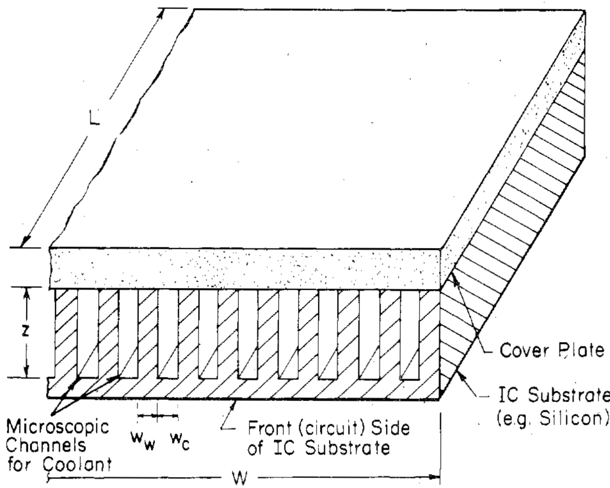

Fig. 1. Schematic view of the compact heat sink incorporated into an integrated circuit chip. For a 1 cm^2 silicon IC using a water coolant, the optimum dimensions are approximately $w_w = w_c$ 57 μm and z = 365 μm.

channels, we have

$$\theta_{\text{conv}} = \frac{2}{k_f \text{Nu}_\infty LW}(w_c \alpha^{-1} \eta^{-1}). \quad (1)$$

We can get an analytical approximation for η by assuming a constant heat-transfer coefficient h up the walls (a good assumption provided η is not too small) and modeling the heat flow in the walls as one-dimensional:

$$\eta = \frac{\tanh N}{N},$$

where

$$N = (2h/k_w w_w)^{1/2}\ z$$

$$= (\text{Nu}_\infty k_f/k_w)^{1/2} \frac{w_c + w_w}{2(w_c w_w)^{1/2}} \alpha. \quad (2)$$

η is thus a monotonically decreasing function of N, with $\eta \simeq 1$ for $N \ll$ and $\eta \cong N^{-1}$ for $N \gg 1$.

As discussed, there will probably be some maximum pressure P available to pump the coolant. The mean flow velocity v in our high-aspect ratio channels can then be calculated, assuming laminar flow between parallel plates: $v = w_c^2 P/12\mu L$. The total volume flow rate is easily seen to be $f = \frac{1}{2} v W w_c \alpha$, whence

$$\theta_{\text{heat}} = \frac{1}{\rho C_p f} = \frac{24\mu L}{\rho C_p P W}(w_c^{-3} \alpha^{-1}). \quad (3)$$

We seek an optimum choice of design variables w_w,w_c, and α which minimizes the total thermal resistance $\theta_{\text{conv}}(w_w, w_c, \alpha) + \theta_{\text{heat}}(w_c, \alpha)$. Referring to equations 1 and 2, we see that for any w_c and α, we can minimize thermal resistance by maximizing η, which means $w_w = w_c$.

Both θ_{conv} and θ_{heat} decrease monotonically with increasing α, so there is no theoretical optimum value for α. However, the fin efficiency η rolls off as α^{-1} for large α, hence θ_{conv} asymptotically approaches a lower limit $\theta_{\min} = w_c/k_{\text{eff}} LW$, where $k_{\text{eff}} = \sqrt{\frac{1}{4}\text{Nu}_\infty k_w k_f}$ can be viewed as an effective thermal conductivity for the heat sink assembly. This result is significant, for it indicates the highest performance (lowest θ) which we can expect to achieve with liquid cooling (within the framework of our model), given the channel width and substrate and coolant thermal conductivities. For a water-cooled silicon substrate with very high-aspect ratio channels, $k_{\text{eff}} = 0.13$ W/°C-cm. The maximum allowable circuit power density is $(\dot{Q}/LW) = (\Delta T_{\max}) k_{\text{eff}}/w_c$, which confirms that microscopically narrow channels are the key to efficient heat removal: if $w_c < 50$ μm and $\Delta T_{\max} = 50$°C, then over 1300 W/cm^2 can be dissipated!

For a practical design, we choose an aspect ratio $\alpha_c = \sqrt{k_w/k_f \text{Nu}_\infty}$, for which $N = 1$, $\eta = 0.76$, and hence

$$\theta_{\text{conv}}\Big|_{\substack{w_w = w_c \\ \alpha = \alpha_c}} = 1.31\theta_{\min} = \frac{1.31}{LWk_{\text{eff}}} w_c. \quad (4)$$

Further increases in alpha would provide only small reductions in θ_{conv} as it approaches $\theta_{\min}$. Referring to equations 3 and 4 and setting $\alpha = \alpha_c$, we see that θ_{heat} varies as w_c^{-3} and θ_{conv} varies as w_c, hence an optimum channel width exists which minimizes their sum, θ:

$$w_c = 2.29\sqrt[4]{\mu k_f L^2 \text{Nu}_\infty/\rho C_p P},$$

for which

$$\theta = \frac{4}{3}\theta_{\text{conv}} = \frac{8.01}{WL^{1/2}}\sqrt[4]{\mu/k_f k_w^2 \rho C_p P \text{Nu}_\infty}.$$

For a water-cooled silicon heat sink on a (1 cm) × (1 cm) substrate, a water pressure of $P = 30$ psi $= 2.07 \times 10^6$ dynes/cm^2, our design procedure gives:

$w_c = w_w = 57$ μm;
$\alpha_c = 6.4$, so $z = 365$ μm, which conveniently is a typical IC silicon wafer thickness!
$\theta = 0.086$°C/W at $f = 11$ cm^3/s.

(We have used $\text{Nu}_\infty = 6$, which is about right for this aspect ratio). Note that $L/(D \cdot \text{Re} \cdot \text{Pr}) = 0.018$ and $\text{Re} = 730$ for our design, so our assumptions were self-consistent (laminar flow and an almost fully-developed temperature profile).

Experiments

Using the preceding parameters as guidelines, we have fabricated and tested several high-performance heat sinks. In a series of experiments, 50-μm wide channels with 50-μm wide walls were etched vertically using KOH (an orientation-dependent etch) [7] to a depth of about 300 μm in ⟨110⟩ silicon wafers of thickness 400μm. A Pyrex cover plate was anodically bonded [9] over the channels and over a pair of etched manifolds at the ends of the channel array. Deionized water at approximately 23°C was fed into the input manifold through a hole in the cover plate at pressures up to 31 psi, and drained from the output manifold through a similar hole. Heat was supplied by a thin-film WSi_2 resistor approximately (1 cm) × (1 cm) in area and 1 μm thick, which was sputtered onto the

Authorized licensed use limited to: Stanford University. Downloaded on May 8, 2009 at 19:52 from IEEE Xplore. Restrictions apply.

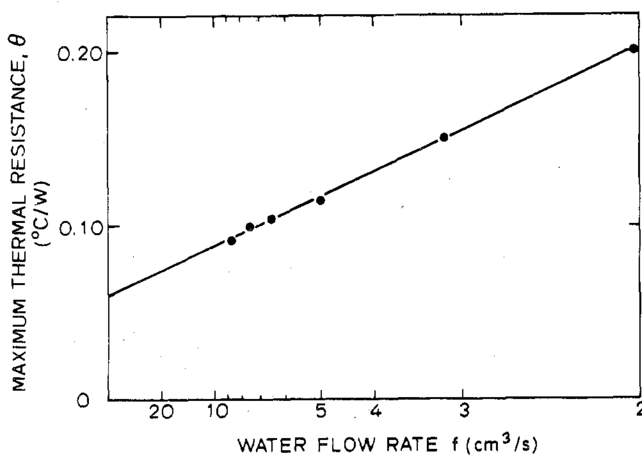

Fig. 2. Measured values of maximum (downstream) thermal resistance Θ as a function of inverse flow rate $1/f$ for heat sink no. 3 of Table 1. As predicted, the data fall on a straight line, implying fully-developed temperature profiles.

thermally-oxidized front surface of the wafer. Thermocouples monitored the temperature of the input and output water and the heater resistor (the latter was measured near the downstream end, where the temperature is highest). We confirmed that the flow rate obeyed Poiseuille's equation and that the thermal resistance was independent of power level. Table 1 summarizes the results obtained for three different heat sinks having similar parameters; all had maximum (downstream) thermal resistances of about 0.1°C/W for a 1 cm^2 area, as expected. One device was tested to 790 W/cm^2.

A further confirmation of the theory was obtained by examining the dependence of the maximum thermal resistance on water flow rate f (cm^3/s). θ_{conv}, the thermal resistance due to conduction from the front of the wafer to the channel region, is clearly independent of f. The same is true for θ_{conv}, provided we have the predicted fully-developed temperature profile. θ_{heat} will be inversely proportional to the flow rate. Thus a plot of $\theta = \theta_{cond} + \theta_{conv} + \theta_{heat}$ vs. f^{-1} should yield a straight line, and experimentally this was indeed the case (Fig.2).

Although a uniform thin-film resistor was used as a heat source in our experiments, in an actual IC the heat is generated in localized areas such as p–n junctions. This will result in an extra contribution to θ due to thermal spreading resistance. This term would be exceedingly small ($\ll 0.01$°C/W) in a VLSI circuit consisting of thousands of uniformly-distributed devices, but it may be important in specialized ICs consisting of only a few localized high-power heat sources.

The dramatic (forty-fold) improvement in practical, compact IC heat-sinking capability presented here offers a new degree of freedom for the system designer. For example, speed-power tradeoffs can now be resolved in favor of more speed, and in particular ECL circuitry may now be a more attractive candidate for high-speed VLSI. The low thermal resistance may also be useful for moderate-power ICs where the temperatures of different components must match closely or be held close to the coolant temperature. The incorporation of this very compact integral heat sink into a conventional IC package is relatively straight forward.

Acknowledgments

We would like to thank K. Bean of Texas Instruments, Inc. and P. Barth, J. Beaudouin, W. Kays, J. Plummer, K. Saraswat, J. Shott, and R. Swanson of Stanford University for their help. One of us (D.B.T.) was supported by the Fannie and John Hertz Foundation. This work was partially supported by the Joint Services Electronics Program.

References

[1] R. K. Watts, W. Fichtner, E. N. Fuls, L. R. Thibault, and R. L. Johnston, "Electron Beam Lithography for Small MOSFETs," *IEDM Technical Digest*, pp. 772–775, 1980.

[2] "Logic Packaging in the IBM 3081," *Electronic News*, p. 47, vol.17 Nov. 1980.

[3] W. Anacker, "Liquid Cooling of Integrated Circuit Chips," *IBM Tech. Disclosure Bulletin*, vol. 20, pp. 3742–3743, 1978.

[4] R. W. Keyes, "Physical Limits in Digital Electronics," *Proc. IEEE*, vol. 63, pp. 740–767, May 1975; "Fundamental Limits in Digital Information Processing," *Proc. IEEE*, vol. 69, pp. 267–278, Feb. 1981.

[5] C. Y. Ho, R. W. Powell, and P. E. Liley, *J. Phys. Chem. Ref. Data*, vol. 3, Suppl. 1, I-588, 1974.

[6] W. M. Kays and M. E. Crawford, *Convective Heat and Mass Transfer*. New York: McGraw-Hill, 1980, ch. 8.

[7] W. M. Kays and A. L. London, *Compact Heat Exchangers*, 2nd ed. New York: McGraw-Hill, 1964, p. 14.

[8] K. Bean, "Anisotropic Etching of Silicon," *IEEE Trans. Electron Devices*, vol. ED-25, no. 10, pp. 1185–1193, Oct. 1978.

[9] G. Wallis and D. I. Pomerantz, "Field-Assisted Glass-Metal Sealing," *J. Appl. Phys.*, vol. 40, no. 10, Oct. 1969, pp. 3946–3949.

Authorized licensed use limited to: Stanford University. Downloaded on May 8, 2009 at 19:52 from IEEE Xplore. Restrictions apply.