UCRL-53515

Heat-Transfer Microstructures for Integrated Circuits

David Bazeley Tuckerman
(Ph.D. Thesis)

February 1984

Lawrence Livermore National Laboratory

DISCLAIMER

Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Accession Number: 3181

Publication Date: Feb 01, 1984

Title: Heat-Transfer Microstructures for Integrated Circuits

Personal Author: Tuckerman, D.B.

Corporate Author Or Publisher: Lawrence Livermore National Laboratory, Univ. of Cal., Livermore, CA 9 Report Number: UCRL-53515

Comments on Document: Dissertation in partial fulfillment of requirements for degree of Doctor of Philosophy at Stanford University Inventory for TN

Descriptors, Keywords: Heat-Transfer Microstructure Integrated Circuit

Pages: 00141

Cataloged Date: Oct 11, 1991

Contract Number: W-7405-ENG-48

Document Type: HC

Number of Copies In Library: 000001

Record ID: 22667

UCRL-53515
Distribution Category UC-38

Heat-Transfer Microstructures for Integrated Circuits

David Bazeley Tuckerman

(Ph.D. Thesis)

Manuscript date: February 1984

LAWRENCE LIVERMORE NATIONAL LABORATORY
University of California • Livermore, California • 94550

Available from: National Technical Information Service • U.S. Department of Commerce
5285 Port Royal Road • Springfield, VA 22161 • $16.00 per copy • (Microfiche $4.50)

Heat-Transfer Microstructures for Integrated Circuits

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

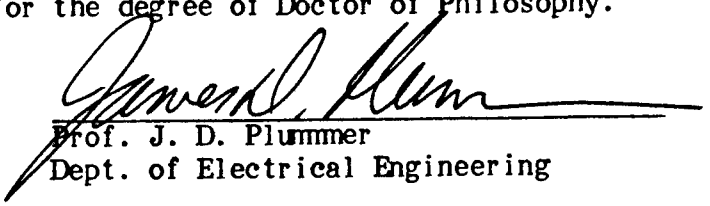
By
David Bazeley Tuckerman

February 1984

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Prof. R. F. W. Pease
Principal Advisor
Dept. of Electrical Engineering

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Prof. J. D. Plummer
Dept. of Electrical Engineering

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Prof. R. J. Moffat
Dept. of Mechanical Engineering

Approved for the University Committee
on Graduate Studies:

Dean of Graduate Studies & Research

Abstract

The design of high-speed integrated circuits and systems is often constrained by thermal considerations. As late as 1981, it was authoritatively predicted that the maximum achievable power flux for liquid-cooled, densely-packed integrated circuits (ICs) would be about 20 W/cm^2.

Convective heat-transfer theory indicates that well over 1000 W/cm^2 can be compactly removed from ICs at normal operating temperatures, provided microscopic (e.g., 50-μm wide) extended-surface structures are used. The difficulty of constructing high-conductance, low-stress thermal interfaces between ICs and heat sinks suggests the use of an integral heat sink. Accordingly, IC microfabrication techniques were employed to design, fabricate, and test novel, ultracompact water-cooled, laminar-flow, optimized plate-fin and pin-fin heat sinks directly within standard-thickness silicon substrates. Worst-case thermal resistances as low as 0.083°C/W were measured from 1-cm^2 thin-film resistors (e.g., a 108°C temperature rise at 1309 W), in good agreement with predictions. Further increases in heat transfer are achievable.

The use of integral liquid-cooled heat sinks in multichip systems presents potential yield, reliability, cost and packaging problems. Attachment of unmodified ICs to micro-heat sinks seems a more attractive approach. A novel die-attachment technique has been developed which avoids the problems of conventional attachments. In this technique, a liquid partially fills an array of micron-wide reentrant capillaries in the heat sink substrate, so that surface tension holds the polished back of an IC in intimate thermal contact with the heat sink. The bond is void-free, virtually stress-free, long-lived, and allows repeated detachment and replacement of ICs without damaging the heat sink substrate. The reentrant grooves were fabricated by a novel process using electroless plating of nickel onto vertical silicon microgrooves. For a 1-cm^2 area, typical interfacial thermal resistances of 0.022°C/W at 300 W have been measured.

In summary, microfabrication techniques have been employed to fabricate new, very high-performance liquid-cooled heat sinks having negligible volume (0.1 cm^3), and also to make a novel, stress-free, reusable microcapillary thermal interface between such heat sinks and integrated circuit substrates. These techniques allow the VLSI system designer more freedom, in that power consumption may be greatly increased while simultaneously realizing the reliability benefits of a much lower operating temperature.

Acknowledgments

First and foremost, I would like to thank my thesis supervisor Prof. R. Fabian Pease for his enthusiastic supervision and support of this work. I would also like to express my appreciation to the Fannie and John Hertz Foundation which provided the financial support for my graduate education and hence made possible a relatively unconstrained choice of research project at Stanford University. The encouragement of Lowell L. Wood of Lawrence Livermore National Laboratory (Hertz Foundation Graduate Fellowship Coordinator) throughout this work was greatly appreciated.

A number of individuals at Stanford provided invaluable consultations during this work. Profs. R. J. Moffat, W. M. Kays and A. L. London (masters of the art and science of heat transfer) were extremely helpful. Profs. J. D. Plummer, T. W. Sigmon, and R. M. Swanson were valuable instructors on various aspects of semiconductor technology, as were P. Barth, J. Beaudoin, J. McVittie, and J. Shott. I would also like to acknowledge discussions with R. W. Keyes of IBM Corp. and B. H. Whalen of TRW Corp.

Valuable supplies of materials and equipment were provided by K. Bean of Texas Instruments Inc., Pei-Yu Wu of Stanford University, and B. McWilliams of Lawrence Livermore National Laboratory.

The technical assistance of W. Holmes and Z. Norris of Stanford University is gratefully acknowledged.

Thanks are due Profs. J. D. Plummer and R. J. Moffat for reading the manuscript.

The final preparation of this document was done while the author was a staff member at Lawrence Livermore National Laboratory.

Financial support for various parts of this research was provided by the Semiconductor Research Corporation, TRW Corporation, Honeywell Corporation, and the Joint Services Electronics Program.

Table of Contents

List of Figures

List of Tables

List of Symbols

Note: The metric units (e.g. Watts, gm, cm, ^{o}C, dynes/cm^2) used in this work are not all standard SI units, but they are traditionally used in the semiconductor industry.

English letter symbols

A = silicon substrate area, cm^2;
also, load-bearing area of an entrapped dust particle, cm^2

A_f = fin area, cm^2

c_f = local friction coefficient or "friction factor" of channels

c_{fm} = mean friction coefficient

C = heat capacity at constant pressure, J/gm-^{o}C

C_v = flow coefficient (e.g., of a filter)

d = effective molecular collision diameter, cm

D = channel hydraulic diameter = $4 \cdot$(cross-sectional area)/(perimeter), cm;
also, elastic plate-deflection modulus $D = Et^3/[12(1-\nu^2)]$, dynes-cm;
also, diffusivity of chemical species, cm^2/sec

E = elastic (Young's) modulus, dynes/cm^2

f = coolant volume flow rate, cm^3/sec

F_{dust} = force supported by a trapped dust particle between plates, dynes

G = shear modulus, dynes/cm^2

h = convective heat-transfer coefficient, W/cm^2-K;
also, enthalpy of coolant, J/gm

h_{IC} = effective heat-transfer coefficient from surface of IC, W/cm^2-K

H = fin height (in z-direction), cm

$H_s = t_{Si} - H$, residual substrate thickness, cm

I = current, amperes

j_y = heat flux into coolant from fins (in y direction), W/cm^2

$j_H = Nu/(Re \cdot Pr^{1/3})$, the Colburn factor

J_z = heat flux up fins (in z direction), W/cm^2

J_S = sheet current, A/cm

k = thermal conductivity, W/cm-K

k_B = Boltzmann's constant, 8.617×10^{-5} eV/K

k_c = coolant thermal conductivity, W/cm-K

k_{avg} = average thermal conductivity of silicon/coolant aggregate (for thermal spreading calculations), W/cm-K

$k_{eff} = \sqrt{k_c k_w} \cdot 2\sqrt{w_c w_w}/(w_c + w_w)$, W/cm-K

k_s = thermal conductivity of substrate ("wall" or "fin"), usually silicon, W/cm-K

k_B = Boltzmann's constant, 8.617×10^{-12} eV/K

K = loss coefficient [pressure drop = $K \cdot (\rho v^2/2)$]

K_c = contraction loss coefficient (entering a header)

K_e = expansion loss coefficient (exiting a header)

K_∞ = loss coefficient due to development of momentum boundary layer

L = length of heated area (sometimes equal to L_s), cm

L^* = dimensionless channel length, $L/(D \cdot Re \cdot Pr)$

L_s = length of cooling channels (i.e., substrate length), cm

L_o = decay length of lateral thermal spreading in silicon heat sink, cm

L_H = length of header (in direction of flow), cm

m = normalized fin height (number of characteristic lengths, seldom exceeds 2); also, molecular mass, gm

n = number of channels in heat sink

Nu = Nusselt number, hD/k_c

Nu_x = local Nusselt number at location x

Nu_L = local Nusselt number at location L (output end of channel)

Nu_∞ = Nusselt number for fully-developed temperature profile

p = perimeter of cooling channel, cm

P = pressure drop in channels, dynes/cm^2

P_o = suction pressure in microcapillary interface, dynes/cm^2

P_{core} = pressure drop in channels neglecting all entrance and header effects, dynes/cm^2

P_∞ = stagnation pressure, dynes/cm^2

Pr = Prandtl number = $\mu C/k$

$\dot{q}$ = heat flow (total heat-transfer rate), W

$\dot{q}''$ = heat flux (heat-transfer rate per unit of area), W/cm^2

$\dot{Q}$ = mechanical pumping power, W

$\dot{Q}''$ = mechanical pumping power **per unit area of cooled area**, W/cm^2

$\dot{Q}''_{crit}$ = mechanical pumping power per unit area below which laminar flow designs are believed superior (conservative estimate), W/cm^2

r = radius or radial distance, cm

R = $LW\theta$, thermal resistance normalized to unit area, cm^2-°C/W

R_{cal} = $LW\theta_{cal}$, cm^2-°C/W

R_{cond} = $LW\theta_{cond}$, cm^2-°C/W

R_{conv} = $LW\theta_{conv}$, cm^2-°C/W

R_{int} = $LW\theta_{interface}$, cm^2-°C/W

R_{opt} = $LW\theta_{opt}$, cm^2-°C/W

Re = Reynolds number = $vD\rho/\mu$

t = thickness (of silicon, solder, oil, etc. as subscripted), cm; also, time, sec

T = temperature (°C, °F, or K, as specified)

T_A = ambient temperature (refers either to ambient air or to cooling water, depending on context)

T_{Debye} = Debye temperature
T_c = mixed mean fluid (coolant) temperature, °C or K
T_j = semiconductor device (junction) temperature, °C or K
T_w = wall (fin) temperature, °C or K
$u_{-1}(x)$ = Heaviside step function
v = velocity (usually a mean fluid velocity unless $\underline{v}$ is used, in which case it refers to local velocity), cm/sec
v_o = free stream velocity, cm/sec
$\underline{v}$ = mean velocity, cm/sec
$v_x(y)$ = local velocity at position y **in x-direction**, cm/sec
V = voltage, V
$\dot{w}$ = mass flow, gm/sec
$w(r)$ = vertical deflection of an elastic plate, cm
w_{bow} = interfacial gap due to wafer warpage or "bow", cm
w_d = dust particle elevation, cm
w_c = channel width, cm
w_m = microcapillary groove width at meniscus level, cm
w_w = wall width, cm
W = width of cooled area (perpendicular to direction of flow), cm
x = distance from input end of channels for flow-friction calculations or from, or start of heated area for heat-transfer calculations (caloric heat-transfer axis), cm
x^* = dimensionless axial distance along channel, $x/(D \cdot Re \cdot Pr)$
x^+ = hydrodynamic entry length, cm
x_{th}^+ = thermal entry length, cm
y = distance from fin surface into coolant (convective heat-transfer axis), cm
z = distance up from base of fin (conductive heat-transfer axis), cm

Greek letter symbols

α = channel surface area enhancement = (channel surface area)/(heated area)
α^* = channel aspect ratio, w_c/H
α_c = "characteristic" value of α; the maximum increase in effective heat-transfer area
β = miscellaneous constant
γ = surface tension, ergs/cm^2, dynes/cm
δ = thickness of a momentum boundary layer, cm
δ_{th} = thickness of a thermal boundary layer, cm
Δ = designates a difference when used as a prefix
$\varepsilon = \alpha/\alpha_c$
ε_{th} = thermal strain
η = combined temperature effectiveness of finned and prime surfaces
η_f = temperature fin effectiveness factor (finned area)
θ = thermal resistance, °C/W
θ_{bulk} = thermal resistance of bulk substrate (e.g., silicon), °C/W
θ_{cal} = caloric thermal resistance due to coolant heating, °C/W

θ_{conv} = convective thermal resistance, °C/W
θ_H = optimized thermal resistance for a fixed fin height H, °C/W
θ_∞ = optimized thermal resistance for an infinite fin height, °C/W
θ_{int} = thermal resistance across an interface, °C/W
θ_{opt} = optimized thermal resistance, °C/W
θ_{spread} = thermal spreading resistance, °C/W
$\Theta_c(x,z)$ = local coolant temperature averaged in y-direction but not in x-direction, °C
κ = thermal diffusivity, $k/\rho C$, cm^2/sec
λ = characteristic heat penetration length up a fin, cm
μ = dynamic viscosity, gm/cm-sec (poise)
ν = kinematic viscosity, cm^2/sec (stokes);
also, Poisson's ratio ($\nu = 0.09$ for silicon)
ρ = density, gm/cm^3;
also, radius of curvature, cm;
also, normalized distance from a dust particle ($\rho = r/R$)
σ = tensile stress, $dynes/cm^2$
σ_{rad} = Stefan-Boltzmann constant, 5.679×10^{-12} $W/cm^2\text{-}K^4$
σ_o = mean sheet conductivity, $(\Omega/\square)^{-1}$
$\sigma_s(x,y)$ = local sheet conductivity, $(\Omega/\square)^{-1}$
σ_{Si} = fracture stress of silicon, 3.47×10^9 $dynes/cm^2$
τ = shear stress, $dynes/cm^2$
τ_d = characteristic reaction time for a monolayer of species to deposit
Υ = dimensionless thermal resistance, $k_{eff}LW\theta/D$
φ = potential, volts
Φ = $c_{f\infty}Re$ (product of fully-developed friction factor and Reynolds number)
Φ_χ = c_fRe (product of Reynolds number and friction factor at given $\chi = D \cdot Re/L$
χ = $D \cdot Re/L$,the reciprocal of the Peclet number
ω = angular frequency, radians/sec

subscripts

c = coolant
lam = laminar
m = mean
o, opt = optimum design point
s = refers to heat sink substrate (e.g. silicon)
th = thermal (as opposed to hydrodynamic)
turb = turbulent
w = wall (i.e., fin material)
∞ = asymptotic (long-channel) limiting value

Glossary

chip, die, wafer	Used interchangeably in this work to mean a slice of silicon containing integrated circuits. Unless specified, the size of the slice is arbitrary.
CFOM	Coolant Figure of Merit
CMOS	Complementary Metal-Oxide-Semiconductor
ECL	Emitter-Coupled Logic
IC	Integrated Circuit
MOS	Metal-Oxide-Semiconductor
MTF	Median Time to Failure
ODE	Orientation-Dependent Etching
RMS	Root Mean Square
SEM	Scanning Electron Micrograph or Scanning Electron Microscope
TCE	Temperature Coefficient of Expansion
TCR	Temperature Coefficient of Resistance
VLSI	Very Large-Scale Integrated or Very Large-Scale Integration

Chapter 1
Introduction and Background

1.1. Preface

The design of modern computing systems comprised of millions of switching elements calls upon a large range of engineering skills and talents. A common and convenient partitioning has evolved between system architects and device designers. Device designers and device physicists concern themselves primarily with the detailed physics of individual devices, in order to produce an optimized logic family. System architects usually deal with the logic gate as their smallest building block; the gate is characterized by its function and propagation delay, but all the physics has been removed from the model. However, in real systems a variety of physical and technological problems arise from the **aggregation** of large numbers of devices into a system. These system-induced problems are not necessarily anticipated by the device physicist, as they may not arise at small scales of integration. They likely will not be anticipated by the systems architect, because the relevant physical phenomena do not exist at the level of abstraction of a logic gate. Perhaps we should categorize those who work on such problems as "system physicists" (as opposed to device physicists).

One example of a "system physics" problem is system reliability. A single logic device having a median time to failure (MTF) of 100 years may, for practical purposes, be considered perfectly reliable. Yet a system fabricated out of millions of such devices would have an unacceptably short lifetime, assuming a typical failure distribution function and ignoring the possibility of a fault-tolerant design. The physical mechanisms responsible for "hard" failures might be electromigration [1], intermetallic formation, impurity diffusion through dielectrics [2], etc.; "soft" errors could also occur, for example due to alpha particles [3]. The problem of integrated circuit (IC) manufacturing yield might also be considered a system physics problem, and physical phenomena which influence yield continue to be uncovered.

Another system physics problem involves unwanted interactions among nominally independent devices; for example, electromagnetic coupling (crosstalk) between adjacent long, parallel wires on an IC, especially when a ground plane is remote or absent [4]. The

phenomenon of "sneak" paths in memory cells due to parasitic capacitances is another well known example [5]. Often the interaction may not be a problem except when a number of logic gates switch simultaneously; this can cause a power-supply voltage fluctuation due to the presence of a relatively large parasitic resistance or inductance in series with the power bus [6].

Interconnections (wiring) between circuits, both intra-chip and inter-chip, are another system physics problem. The design of suitably compact, low-inductance, low-resistance, low-crosstalk, high-speed inter-chip connectors for high-performance computers has long been a problem to the packaging engineer. More recently, it has become clear that the design of high-speed intra-chip wiring is a formidable problem as critical dimensions are shrunk [7, 8].

Yet another system physics problem is the removal of the heat generated by densely-packed arrays of integrated circuits. That is the subject of this thesis, and much more will be said of it in the following chapters.

The purpose of this preface is to convince the reader of the importance of studying the physical problems associated with large electronic systems. Historically these problems have been addressed whenever they present a major (often unexpected) impediment to system design, but not otherwise. Thus the progress which is made tends to be very incremental, adequately solving the problem at hand but not anticipating future needs. As a consequence, the device and system architects tend to accept the constraints of system physics as fairly firm and not subject to drastic improvements. However, it is a tenet of this thesis that major efforts in improving heat transfer, reliability, yield, device interactions, interconnect technology, and packaging would in many cases benefit system performance far more than would an equivalent amount of time and money spent on improving device design or system architecture.

An important physical problem in computer systems is that of heat dissipation and removal. The heat generated by a single logic gate is small: at most a few milliwatts for most high-speed logic such as ECL; much less for other logic families such as CMOS (Fig. 1-1). When tens or hundreds of thousands of logic gates are fabricated on a single VLSI chip, the total power consumption could conceivably reach the kilowatt level if high-performance logic is used. Even if a low-power logic family is used, a complete high-performance computer system might have 10^7 to 10^8 switching elements (e.g., transistors) and hence comprise

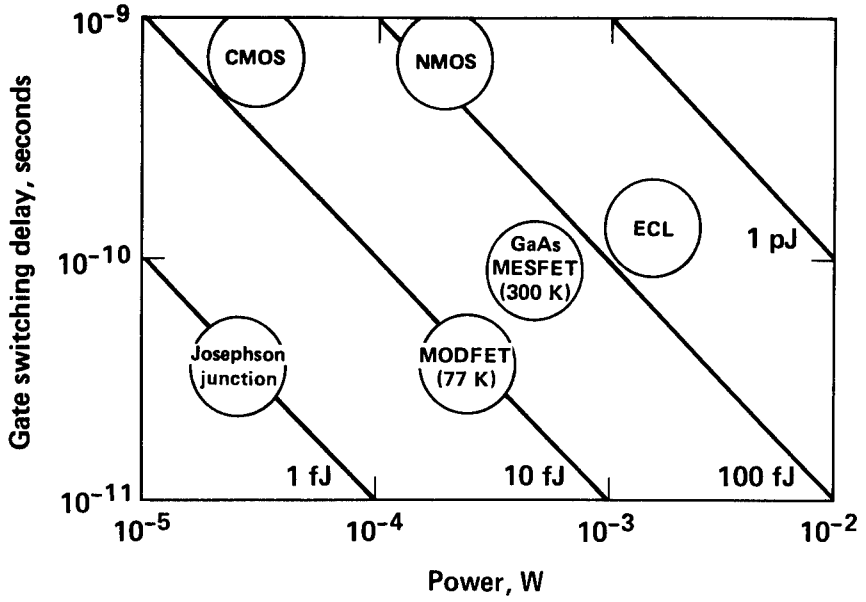


Figure 1-1: Speed-power relationships for various logic families (adapted from Ref. [9]).

hundreds of chips. If these chips are packed very closely together to minimize propagation delays, the problem of removing tens (or even hundreds) of kilowatts of heat while maintaining normal circuit temperatures (usually less than 120°C; preferably even lower for enhanced reliability) from a system volume of less than 1 liter becomes challenging.

It has been widely claimed that such a heat-removal task is effectively impossible. One pioneer of system physics has estimated that forced-air cooling of logic chips is limited to power densities of about 1 W/cm^2, and that liquid cooling is limited to about 20 W/cm^2 [10, 11]. Another leader in high-speed Josephson systems has stated that with present technology it would be impossible to remove 20 kW from a room-temperature computer having a volume less than 640 cm^3, and that even a tenfold reduction in power to 2 kW would still present a "difficult, if not impossible" cooling task [12]. Indeed, the best commercial technologies presently available for cooling densely packed arrays of ICs are the IBM Thermal Conduction Module (TCM) [13, 14] and the Honeywell Silent Liquid Integral Cooler (SLIC) [15], both of which are limited to heat fluxes of $\sim$20 W/cm^2. However, the arguments that this represents a practical limit are not convincing, and we shall show that these estimates are low by at least 2 orders of magnitude.

In this work the physical limits on compact cooling of densely packed arrays are critically examined, and two new technologies are described which allow one to approach these limits. The remainder of Chapter 1 reviews the physics of heat transport in materials, and describes

the components of thermal resistance in conventional integrated circuit packaging. Chapter 2 develops the theory and design optimization of ultracompact liquid-cooled laminar-flow heat sinks (particularly water-cooled silicon), using classical convective heat-transfer theory [16] and compact heat-exchanger theory [17]. Various refinements in the design as well as techniques for further improving performance are discussed. The specific optimization procedures and the suggestion of scaling compact heat-exchanger technology to microscopic duct dimensions ($\sim$50 μm) for cooling IC's are believed to be original. The idea of constructing an integral heat sink in silicon chips is not new [12, 18, 19], but previous workers evidently performed no analyses nor experiments, hence clearly did not appreciate the extraordinary performance benefits which could result from an optimized micro-heat sink design.

Chapter 3 describes the fabrication and testing of integral silicon micro-heat sinks using variations on conventional IC fabrication technology. The dimensions are about an order of magnitude smaller than in conventional compact heat-exchangers [17, 20]. Accurate flow-friction and heat-transfer measurements were performed. The experiments confirmed the very high performance of the silicon heat sinks, with power levels as high as 1309 W/cm^2 demonstrated at normal operating temperatures. Various portions of the work in Chaps. 2 and 3 have been previously published by the author in Refs. [21], [22], and [23].

Chapter 4 describes the design of a novel, stress-free, detachable thermal interface between ICs and separate micro-heat sinks. The interface uses a liquid as the thermal interface material. The technology would be useful for multi-chip or wafer-scale packaging and/or testing in applications where an integral heat sink is deemed impractical. While liquid thermal interfaces have been suggested previously [24, 25, 26, 27], this configuration is unique in that specially designed microscopic (2-μm wide) partially-filled capillaries were used to insure that the interface is void-free and tolerant of mechanical perturbations (flexure), and to maintain a strong, continual suction pressure between the surfaces to correct for wafer warpage.

Chapter 5 describes the fabrication and testing of the microcapillary thermal interface technology devised in Chapter 4. A novel process was developed for fabricating micron-wide reentrant grooves using electroless nickel plating. Very good thermal performance was demonstrated (typical 0.022°C/W thermal resistance for a 1-cm^2 area). Portions of this work have been published previously in Ref. [28].

Chapter 6 briefly summarizes the results and contributions of this work and suggests directions for future research.

1.2. Physics of Heat Conduction

Heat may be transported by conduction, convection, or radiation; however convection is simply conduction in a moving fluid, so the essential physics is still a conduction process. Radiative heat transport is a negligible factor in cooling of semiconductor devices, because junction temperatures T_j typically must not exceed 120°C. Thus the maximum possible radiated power is

$$\dot{q}'' \leq \sigma_{rad} T_j^4 = (5.679 \times 10^{-12}\ W/cm^2\text{-}K^4) \cdot (393\ K)^4 = 0.14\ W/cm^2 ,$$

which is a rather small heat flux.

Over distances which are long compared to the mean free path of the dominant heat carrier, heat conduction is described by Fourier's Law and by energy conservation:

$$\dot{\mathbf{q}}'' = -k \nabla T$$

$$\rho C (\partial T/\partial t) = -\nabla \cdot \dot{\mathbf{q}}'' ,$$

where k = (thermal conductivity) and ρC = (volumetric heat capacity). If the heat carriers have a constant velocity v and mean free path λ, then from elementary kinetic theory [29] one finds that $k = \rho C v \lambda / 3$. More generally v is not constant, and we would approximate $k \simeq \rho C \underline{v} \lambda / 3$ where $\underline{v}$ is an average carrier velocity. Table 1-1 summarizes typical theoretical and experimental values of these parameters for various forms of matter. Although these numbers are highly approximate, they provide insight into why different materials have the thermal conductivities that they do. Table 1-2 lists room-temperature thermal conductivities of some frequently-encountered materials.

For a perfect monatomic gas, the heat carrier is the gas molecule itself, and kinetic theory predicts that

$$k = (0.54\, k_B^{3/2} T^{1/2}) m^{-1/2} d^{-2} ,$$

where m is the molecular mass, d is the effective molecular collision diameter, and k_B is Boltzmann's constant. The main point is that light (hence fast) molecules with small collision cross sections have the highest thermal conductivity. Thus helium gas has 5.6 times the thermal conductivity of air.

The high thermal conductivity of crystalline metals can be understood in view of the electron's very high velocity ($\sim 10^8$ cm/sec) due to its light mass, and the its long mean free

Type of Material	ρC	$\underline{v}$	λ
Gases	$(3/2)k_B n$ (monatomic)	$\sqrt{8k_B T/\pi m}$ ($\sim 10^5$ cm/sec)	$(\sqrt{2}\pi d^2 n)^{-1}$ (60-300 nm, at STP)
Metals (crystalline)	$\rho C_{el} \simeq 4.9(T/T_F)k_B n$ ($T_F \simeq 2\text{-}17 \times 10^4$ K)	$\underline{v}_{el} = \hbar k_F / m_e$ ($1\text{-}2 \times 10^8$ cm/sec)	$\underline{v}_{el} \tau_{el\text{-}ph}$ (5-100 nm)
Insulators (crystalline)	$3k_B n$	$\underline{v}_{ph} = c_{sound}$ ($\sim 5 \times 10^5$ cm/sec)	$\underline{v}_{ph} \tau_{ph\text{-}ph}$ (2-50 nm)
Liquids or Amorphous Solids	varies ($\sim k_B n$)	c_{sound} ($\sim 10^5$ cm/sec)	intermolecular dist. (0.5 nm)

Notes: T_F = Fermi temperature, c_{sound} = speed of sound, n = # atoms/cm^3,
τ = mean collision time with phonons, m_e = mass of electron.

Table 1-1: Typical room-temperature thermal transport parameters (very approximate).

path. The high thermal conductivity of crystalline insulators such as silicon (thermal conductivity 1/3 that of copper) or diamond (5 times that of copper) at room temperatures is due to the relatively long phonon mean free paths, comparable in magnitude to electron mean free paths in crystalline metals. Although the phonon velocity (speed of sound) is 2 or 3 orders of magnitude less than electron velocities, the phonon (lattice) heat capacity of an insulator is much greater than the electronic heat capacity of a metal, at room temperature. Note that amorphous substances (e.g., liquids or glasses) have thermal conductivities 2 or 3 orders of magnitude less than their crystalline counterparts; this is due to the reduction of the electron or phonon mean free path to intermolecular distances.

The thermal conductivity of silicon will be of particular interest to us; Fig. 1-2 plots it as a function of temperature. The rapid rise in thermal conductivity with decreasing temperature is due to the dramatic increase in phonon mean-free-path length (the phonon population diminishes exponentially with temperature when $T \ll T_{Debye}$). As shown in Fig. 1-3, silicon's room-temperature thermal conductivity is reduced somewhat by doping with impurity concentrations greater than 10^{18} cm^{-3}. The average phonon mean free path λ at 25°C may be approximated from Eq. 1.1 as

	Material	k (W/cm-K)
Gases:	Air	.00026
	Helium	.00145
Liquids:	Water	.0061
	Silicone oil (typ.)	.0015
	Nitrogen @ 77 K	.00140
	Mercury	.0830
Nonmetallic Solids:	Pyrex	0.0109
	Carbon (amorphous)	0.0159
	Carbon (diamond IIA)	23.2
	Silicon (intrinsic)	1.48
	GaAs	0.54
Metallic Solids:	Copper	4.01
	Aluminum	2.35
	Stainless Steel	0.18

Table 1-2: Thermal conductivity of various materials; T = 20-27°C unless noted.

$\lambda = 3k/\rho C\underline{v} = 3(1.48\ W/cm\text{-}K)/(1.64\ J/cm^3\text{-}K)(9.0\times10^5\ cm/sec) = 300$ Angstroms.

Since this is much smaller than any device structures of present interest, we shall use Fourier's law without any corrections from the microscopic theory of heat flow. It should be noted that direct measurement of phonon mean free path lengths by means of transmited phonon drag have found much longer lengths (several microns at room temperature) [32]. However, these phonons are presumably very long-wavelength ones, and due to the low density of states they do not make a major contribution to the heat conduction process.

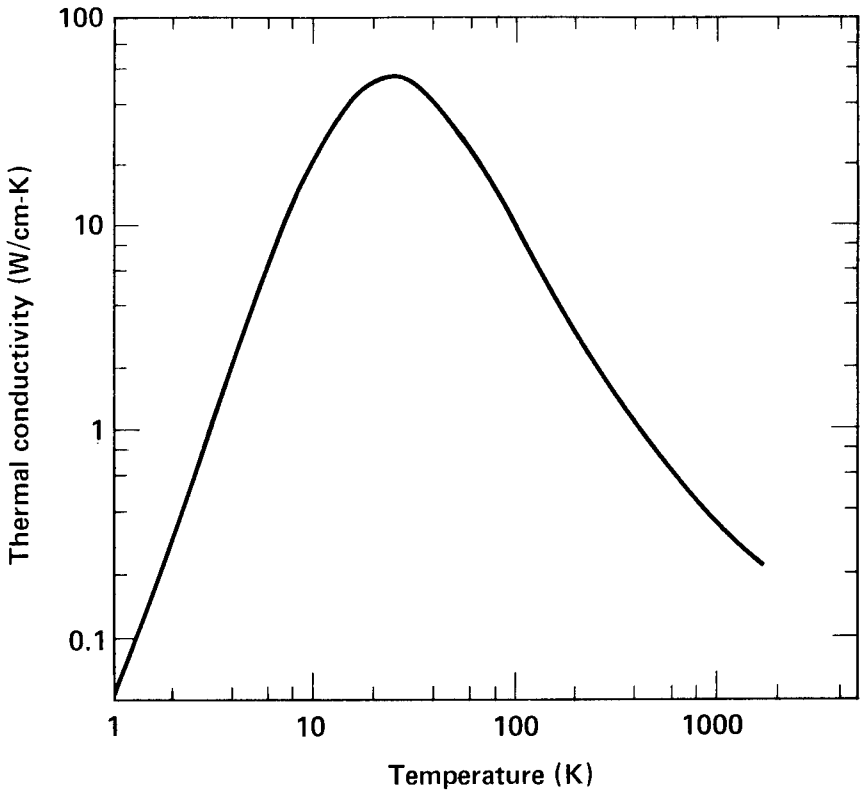


Figure 1-2: Thermal conductivity of intrinsic silicon vs. temperature (from Ho [30]).

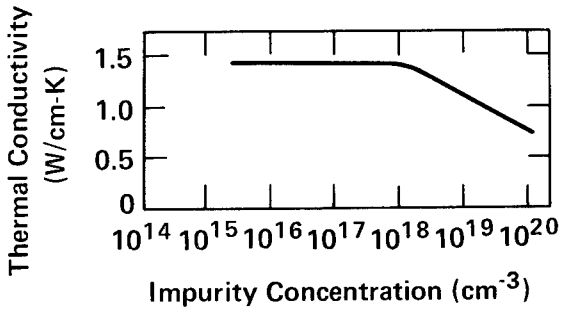


Figure 1-3: Thermal conductivity of single-crystal Si vs. impurity concentration at 300 K. Silicon was doped with As, Sb, P, and Ga (data of Arasli and Aliev [31]).

1.3. Components of Thermal Resistance

The performance of a convective heat-transfer system for integrated circuits is measured by its thermal resistance $\theta = \Delta T/\dot{q}$, where ΔT is the temperature rise of the circuit above ambient, and $\dot{q}$ is the IC power dissipation. This formalism is primarily useful when the heat transfer is a linear problem, i.e., θ is independent of temperature. This is often a good approximation in forced convection cooling systems. For cooling semiconductor integrated circuits, θ is in general the sum of 5 components (Fig. 1-4): θ_{spread}, the spreading resistance from the individual heat-generating devices in the semiconductor substrate; θ_{bulk}, due to heat conduction through the bulk semiconductor; $\theta_{interface}$, the thermal resistance associated with the IC/heat-sink interface (if any); θ_{conv}, the convective thermal resistance between the heat sink and the coolant fluid; and θ_{cal}, the "caloric" thermal resistance due to the heating of the fluid as it absorbs energy passing through the heat sink. We now examine these components to determine which of them present the most fundamental impediments to the compact removal of high heat flux.

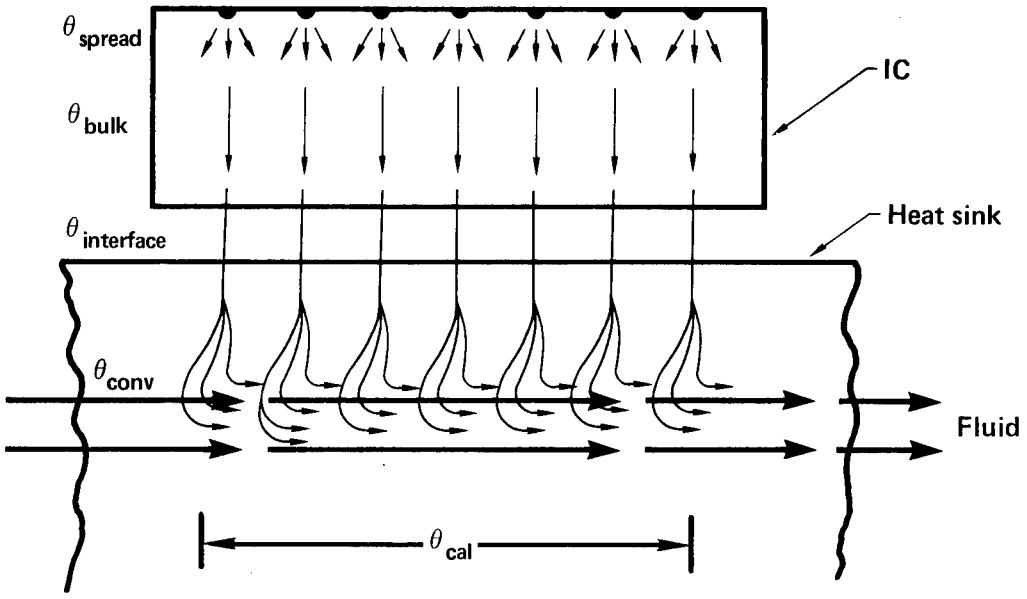


Figure 1-4: Components of thermal resistance in convectively cooled ICs.

Of these 5 components of thermal resistance, θ_{bulk} is the easiest to minimize. For silicon substrates, $\theta_{bulk} = t_{Si}/k_{Si}A$ where t_{Si} = (silicon substrate thickness) and A = (substrate area). Thus if a 1-cm^2 IC substrate is thinned to 100 μm, then θ_{bulk} = .007°C/W at room temperature, which is very small; furthermore there is no fundamental limit to thinning the

substrate even further (down to the thickness of the active device layers, which seldom exceed 20 μm even in bipolar technology).

The caloric thermal resistance θ_{cal} also has a very simple form: $\theta_{cal} = 1/\rho Cf$ where ρC is the volumetric heat capacity and f is the volume flow rate. Obviously θ_{cal} is made small by choosing a coolant fluid having high volumetric heat capacity (water is one of the best; ρC = 4.18 J/°C-cm^3) and using a very high flow rate. For example, 10 cm^3/sec of water provides a caloric thermal resistance of only 0.024°C/W. The limits on reducing θ_{cal} will be determined by the maximum fluid flow rate, which in turn is set by the fluid mechanics of the problem (to be discussed in Chapter 2).

Thermal spreading resistance θ_{spread} is sometimes quoted as an important limiting thermal resistance. However, for the high levels of integration (more than 10^5 devices/cm^2) which are contemplated, it is only a minor contributor, and becomes relatively less important as device dimensions are scaled down. The thermal spreading resistance of a plane square (dimensions $L_j \times L_j$) heat source embedded in the surface of a silicon substrate has been calculated [33] to be $\theta_{spread} = 0.56/k_{Si}L_j$, which for a L_j = 2 μm heat source works out to be θ_{spread} = 1.9°C/mW for each device. A somewhat lower figure is obtained if the heat source is hemispherical rather than planar; then $\theta_{spread} = 1/\pi k_{Si}L_j$ = 1.1 °C/mW. (The actual value for a real device is presumably somewhere in between). Now we believe that for **electrical** (not thermal) reasons it is unlikely that VLSI integrated circuits would be designed for power densities greater than 1000 W/cm^2 (present practice is under 20 W/cm^2). Thus for densities exceeding 10^5/cm^2, one must design devices which consume an average of less than 10 mW each; hence the device temperature rise due to thermal spreading would be less than 20°C, which is usually acceptable. The situation improves as devices are scaled down, because in order to maintain the power density at or below 1000 W/cm^2, the device power must scale as L_j^{-2}, whereas the thermal spreading resistance θ_{spread} increases only as L, hence the temperature rise will decrease as L^{-1}. Of course specific devices such as off-chip driver transistors may substantially exceed the average device power level, but these could be dealt with specially (say, by paralleling a number of smaller-area transistors instead of fabricating one large-area transistor).

The remaining components of thermal resistance are the convective (θ_{conv}) and interfacial ($\theta_{interface}$) components. These are the most difficult ones to optimize in practical, compact computer systems, and in fact one or both of them usually dominate the overall thermal resistance. The optimization of these thermal resistances for semiconductor ICs is the main

subject of this work. Note that one can conceive of eliminating θ_{int} entirely by constructing a heat exchanger as an integral part of the heat-generating circuits; this is the approach taken in Chapters 2 and 3. If this is not practical, a variety of approaches exist for making high-performance thermal interfaces between ICs and heat sinks; Chapters 4 and 5 describe a new technique for doing so which avoids many of the limitations of conventional die attachment techniques.

Chapter 2
Microscopic Silicon Heat Sinks: Theory

This chapter describes the design and optimization procedures for microscopic silicon liquid-cooled heat sinks. The first section briefly reviews some aspects of convective heat-transfer theory which will motivate the design of microscopic liquid-cooled heat sinks for compact cooling of semiconductors. Our discussion will be brief and appeal to intuition; a detailed exposition may be found in the book by Kays and Crawford [16]. All subsequent sections assume familiarity with Ref. [16] or equivalent texts.

2.0.1. The Thermal and Hydrodynamic Boundary Layers

When a fluid having a free-stream velocity of v_o flows past a solid surface or "wall" (e.g., a flat plate, or the inside of a pipe), the fluid velocity will be constrained to be zero at the wall (the "no-slip" boundary condition). The velocity therefore varies from zero at the wall to v_o far from the wall, and the region near the wall over which the majority of the transition occurs is called the hydrodynamic boundary layer or the momentum boundary layer. The boundary layer thickness will increase as one moves downstream until its growth is constrained by another nearby wall (e.g., laminar flow in a pipe) or by the mixing action of turbulent eddies (in the case of turbulent flow). Fig. 2-1 is a sketch of the development of a laminar momentum boundary layer in the hydrodynamic entry region of a flow between parallel plates.

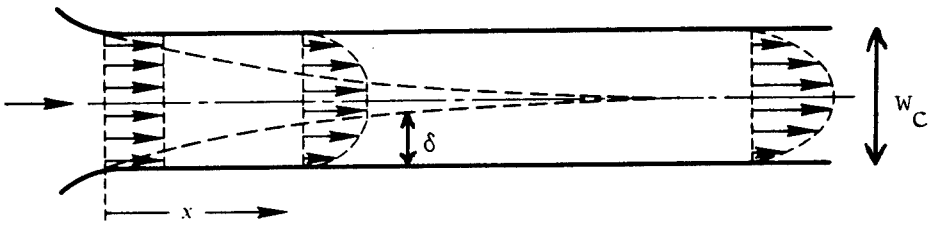


Figure 2-1: Development of a laminar momentum boundary layer between parallel plates.

The approximate functional form of the laminar boundary layer thickness can be deduced by a simple argument. Referring to Fig. 2-1, we define the position along the wall in the

direction of flow as x ($x=0$ at the entrance), and the boundary layer thickness as δ. Momentum transfer in fluids occurs by diffusion, where the diffusion constant is the kinematic viscosity υ (units of cm^2/sec); $\upsilon = \mu/\rho$ where μ is viscosity and ρ is density. Thus the information that the fluid velocity at the wall is zero propagates transversely in the flowing liquid by a diffusion process. In an elapsed time t, this information will be primarily confined within a diffusion length $(\upsilon t)^{1/2}$; this is the approximate boundary layer thickness δ. Since the fluid free-stream velocity is v_o, the fluid outside the boundary layer will have moved a distance $x = v_o t$ in the direction of flow; but since $\delta \simeq (\upsilon t)^{1/2}$, we can eliminate t and we conclude that $\delta \simeq (\upsilon x/v_o)^{1/2}$, i.e., the hydrodynamic boundary layer thickness grows as the **square root** of x. We might consider the velocity profile to be "fully developed" at the position x^+ where this approximate analysis shows δ to be equal to one-half the plate spacing (i.e., the boundary layers merge at the center of the flow). That is, $\delta = w_c/2$, where w_c is the spacing between the parallel walls. Defining the hydraulic diameter D of this parallel-plate pipe in accordance with convention, we have $D = 4 \cdot (\text{cross-sectional area})/(\text{perimeter}) = 2w_c$. Solving for x^+ gives $x^+ \simeq v_o D\rho/16\mu = D \cdot Re/16$, where $Re = v_o D\rho/\mu$ is the Reynolds number. Thus we see that the laminar hydrodynamic entry length should be proportional to the product of the hydraulic diameter D and the Reynolds number Re. Detailed numerical calculations [34] show the actual hydrodynamic entry length x^+ for laminar flow through a pipe to be typically $x^+ \simeq .05(D \cdot Re)$. Here x^+ is defined as the distance beyond which the local friction coefficient (proportional to the shear stress at the wall) is within 2% of its final value.

When a fluid flows through a pipe having a wall temperature T_w which is greater than the mixed mean temperature T_c of the fluid, heat is transported to the fluid. Since the temperature of the fluid in contact with the wall must be continuous (hence equal to T_w), there will develop a thermal boundary layer (in analogy to the momentum boundary layer) over which the major part of the temperature transition from T_w to T_c occurs. Again, the heat is transported by a diffusion process; the diffusion constant is the thermal diffusivity $\kappa = k_c/\rho C$ where k_c is the fluid's thermal conductivity and ρC is its volumetric heat capacity. In a time interval t, the heat will diffuse transversely into the coolant stream a diffusion length $\delta_{th} \simeq (\kappa t)^{1/2}$; this is the approximate thermal boundary layer thickness. If we assume that the velocity profile is laminar and fully developed (parabolic), then the velocity gradient can be approximated as linear near the wall: for parallel plates, $dx/dt = v_x \simeq 12\underline{v}y/D$, where $\underline{v}$ is the mean velocity and y is the distance from the wall. So setting $y = \delta_{th}$ and integrating with respect to time, we get $\delta_{th} \simeq (\kappa D x/8\underline{v})^{1/3}$. Thus we see that the **thermal** boundary layer thickness grows as the **cube root** of x.

As before, the boundary layer might be considered to be fully developed when $\delta_{th} \simeq w_c/2 = D/4$; solving for the thermal entry length gives $x_{th}^{+} \simeq D \cdot Re \cdot Pr/8$ where $Pr = \nu/\kappa$. More precisely, numerical calculations [34] show that the heat transfer typically approaches within 2% of its asymptotic limit when $x \gtrsim .02(D \cdot Re \cdot Pr)$.

To summarize: the laminar momentum boundary layer thickness grows as $x^{1/2}$, and is fully developed in a pipe of hydraulic diameter D when $x \gtrsim .05(D \cdot Re)$. The thermal boundary layer thickness grows as $x^{1/3}$, and is fully developed when $x \gtrsim .02(D \cdot Re \cdot Pr)$. Since $Pr \gg 1$ for nonmetallic liquids, the hydrodynamic boundary layer is fully developed substantially before the thermal boundary layer.

2.1. Elementary Optimization

We now describe the basic optimization procedure which led to the design of microscopic heat sinks. Many approximations and simplifications were made to arrive at the first-order design described here; they will be examined more critically in Section 2.2.

2.1.1. General Considerations

Consider a collection of n identical parallel ducts each of length L (in the x-direction) embedded in a uniformly heated planar substrate of the same length L and width W. A coolant flows in each duct, absorbing a constant heat flow per unit length $\dot{q}/nL$ from its walls (the substrate). The coolant is assumed to be an incompressible Newtonian fluid, i.e., ρ and μ are constants. The use of many separate ducts, rather than a single coolant flow over the entire back of the substrate, allows us to increase the surface area of the structure by an aspect ratio α which is defined as α = (total surface area of ducts)/(area of circuit) = np/W, where p is the cross-sectional perimeter of each duct. We will assume for the moment that at any location x along the length of the duct, the walls are at a uniform temperature $T_w(x)$ around the cross section (i.e., are infinitely thermally conductive). The local convective heat-transfer coefficient h_x is then defined as $h_x = \dot{q}/(nLp[T_w(x) - T_c(x)])$, where $T_c(x)$ is the mixed-mean fluid temperature. Assuming a uniform heat capacity ρC for the coolant, we find from energy conservation that $T_c(x) - T_c(0) = (\dot{q}/\rho CfL)x$ where $T_c(0)$ is the initial coolant temperature which we will denote T_A, and f is the **total** coolant volume flow rate in the ducts. As is customary, we write $h_x = k_c Nu_x/D$, where Nu_x is the local Nusselt number, k_c is the thermal conductivity of the coolant fluid, and $D = 4 \cdot$ (cross-sectional area)/(perimeter p) is the hydraulic diameter of the duct. Thus we have

$$\frac{T_w(x) - T_A}{\dot{q}} = \frac{1}{\rho C f L} x + \frac{1}{\alpha L W k_c} \cdot \left(\frac{D}{Nu_x} \right)$$

The term D/Nu_x can be interpreted as the thermal boundary layer thickness δ_{th} discussed in Section 2.0.1; normally this increases monotonically with increasing x. Thus the maximum wall temperature T_{max} will occur at the output end of the heat sink ($T_{max} = T_w(L)$). The peak thermal resistance θ is therefore:

$$\theta \equiv (T_{max} - T_A)/\dot{q} = 1/\rho C f + D/\alpha L W k_c Nu_L = \theta_{cal} + \theta_{conv} . \quad (2.1)$$

The first term is simply the caloric thermal resistance due to coolant heating, and the second term is the convective thermal resistance.

In order to minimize θ_{conv}, we must minimize the quantity $D/(\alpha Nu_L)$. Traditionally this has been achieved by using high-aspect-ratio ducts to increase surface area (making α large), and by designing to achieve turbulent flow (making Nu large). In order to minimize θ_{cal}, a high coolant flow rate f is desired. The combined requirements of high flow rate and turbulent flow, while maintaining a moderate coolant supply pressure, have led designers to use macroscopic structures such as in water-cooled klystrons [20].

Our approach will be to achieve low convective thermal resistance primarily by minimizing D, i.e., by making the ducts as narrow as possible, rather than by maximizing Nu. This will move the design into the laminar flow regime. The only important lower limit on duct size is set by coolant viscosity. For a given pumping pressure or pumping power, the mass flow rate decreases rapidly as D is reduced, resulting in an increase in θ_{cal}. By assuming a practical limit on the available pressure or power, we can calculate an optimum duct size D which minimizes the overall θ (the sum of θ_{cal} and θ_{conv}). As will be shown later, this optimization procedure results in a thermal resistance which, for the short channel lengths (1 cm) of interest, is comparable to or greater than that which can be achieved with turbulent flow, while yielding a structure which is much more compact (typically by a factor of 20) than conventional turbulent-flow heat sinks [20], for a given operating pressure or pumping power. Since volume is at a premium in high-speed computer systems, we believe that microscopic laminar-flow heat exchangers are necessarily the preferred approach for cooling dense arrays of integrated circuits.

2.1.2. First-Order Design

Fig. 2-2 is a diagram of our high-performance IC heat sink which embodies the principles just discussed.

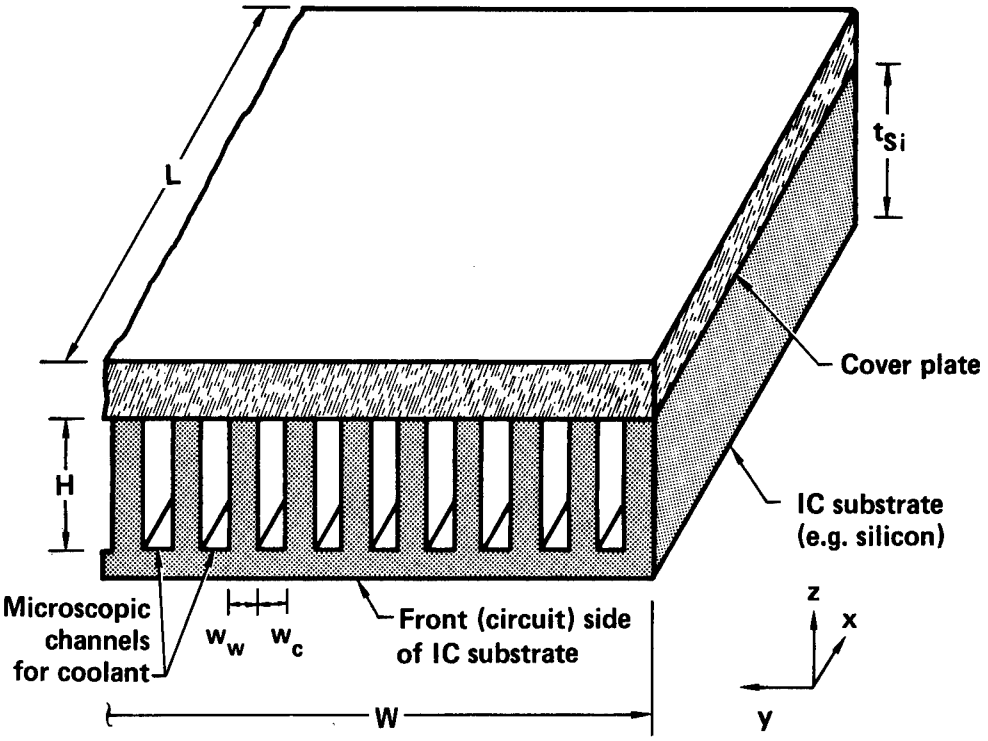


Figure 2-2: Schematic of the compact heat sink incorporated into an IC chip.

The front surface of the substrate (length L, width W, thickness t_{Si}) contains a planar heat source (the circuits) which supplies a spatially uniform heat flux $\dot{q}'' = \dot{q}/LW$, and the back surface contains deep rectangular channels of width w_c and depth H which carry the coolant, separated by walls of width w_w. The surface-area multiplication factor due to the channels is $\alpha = 2(w_c + H)/(w_c + w_w)$, and the hydraulic diameter is $D = 2w_cH/(w_c + H)$. A cover plate is bonded to the back of the substrate to confine the coolant to the channels. In the preceding discussion, we calculated the convective thermal resistance $\theta_{conv} = (1/k_cLW) \cdot (D/\alpha Nu_L)$ by assuming the walls of the channel were infinitely thermally conductive. To account analytically for a finite wall (substrate) conductivity k_w (which implies a nonuniform temperature up the walls), we will have to make some approximations. Assume that the fins have sufficiently high aspect ratio that the heat flow can be modeled as one-dimensional (in the z-direction only). The heat flux J_z up the fin is then defined by the gradient in wall temperature

$$J_z(x,z) = k_w \cdot \partial T_w(x,z)/\partial z \,. \tag{2.2}$$

We shall neglect the heat transferred from the "prime surface" at the bottoms of the channels, and hence assume that all the heat conducts up the "extended surface" fins (Fig. 3-19). This is of course a conservative assumption, and quite reasonable for the 8:1-aspect-ratio ducts which we will be using. The heat flux entering the base of the fin ($z = 0$) is thus

$$J_z(x,0) = \dot{q}''(w_c + w_w)/w_w \tag{2.3}$$

The heat flux at the top of the fins ($z = H$) is assumed to be zero (i.e., the cover plate has a negligibly small thermal conductivity):

$$J_z(x,H) = 0 \tag{2.4}$$

The heat flux j_y from the fin surface into the coolant is determined from energy conservation:

$$2j_y(x,z) = w_w \cdot (\partial J/\partial z) \tag{2.5}$$

j will be driven by the temperature gradient $\partial T/\partial y$ between the wall and the coolant. For our high-aspect-ratio channels the heat transfer at a point on the wall will be sensitive only to the fluid temperature in the vicinity of that point. We define the local mixed mean coolant temperature Θ_c by averaging across the channel width:

$$\Theta_c(x,z) \equiv [\int_0^{w_c} T_c(x,z)v_x(y)\,dy] \,/\, [\int_0^{w_c} v_x(y)\,dy]$$

where $v_x(y)$ is the coolant velocity in the flow direction (assumed to be independent of z for high-aspect-ratio channels). We express the heat flux in terms of a local heat-transfer coefficient h_x (assumed to be constant along the length of the channel),

$$j_y(x,z) = h_x \cdot [T_w(x,z) - \Theta_c(x,z)] \tag{2.6}$$

For high-aspect-ratio channels the hydraulic diameter is $D = 2w_c$, hence $h_x = k_c Nu_x/2w_c$, where Nu_x is the Nusselt number for the constant-heat-flux boundary condition [16]. Finally the heating of the coolant will be determined by energy conservation:

$$w_c \rho Cv(\partial \Theta_c/\partial x) = 2j_y(x,z) \tag{2.7}$$

where

$$v = [\int_0^{w_c} v_x(y)\,dy]/w_c$$

is the average coolant velocity; the total flow rate is $f = nw_c Hv = w_c HvW/(w_c + w_w)$. The coolant temperature is taken to be at ambient initially, which for convenience we designate to be zero:

$$\Theta_c(0,z) = T_A = 0 . \tag{2.8}$$

Despite our simplifying assumptions, we are still unable to find closed-form solutions for T_w and Θ_c which satisfy Eqs. (2.2)-(2.8) exactly. One solution which satisfies all but Eq. (2.6) is

$$T_w(x,z) = \dot{q}'' \cdot \frac{(w_c + w_w)\lambda \cdot \cosh((H-z)/\lambda)}{k_w w_w \cdot \sinh(H/\lambda)} + \Theta_c(x,z) \tag{2.9}$$

where

$$\Theta_c(x,z) = L(w_c + w_w)\dot{q}''/\rho C v w_c H = \dot{q}/\rho C f \quad . \tag{2.10}$$

Here λ is defined by $\lambda^2 = w_w w_c (k_w/k_c Nu_x)$. Eq. (2.6) is satisfied in an integrated sense, i.e.,

$$\int_0^H j_y \, dz = (k_c Nu/2w_c) \cdot \int_0^H (T_w - \Theta_c) \, dz .$$

These approximate solutions (Eqs. (2.9) and (2.10)) were used by this author in Ref. [21]. To see the correspondence with that work, we note that the maximum thermal resistance is

$$\theta = T_w(L,0)/\dot{q} = 1/\rho C f + (2w_c/k_c Nu_L LW) \cdot [(w_c + w_w)/2H] \cdot [(H/\lambda)/\tanh(H/\lambda)]$$

$$= \theta_{cal} + \theta_{conv} . \tag{2.11}$$

Thus

$\theta_{cal} = 1/\rho C f$, and

$$\theta_{conv} = (2w_c/k_c Nu_L LW)\alpha^{-1}\eta_f^{-1} , \tag{2.12}$$

where η_f is the conventional [35] "temperature fin effectiveness factor"

$$\eta_f = (\tanh H/\lambda)/(H/\lambda) , \tag{2.13}$$

and α is the area multiplication factor $2H/(w_c + w_w)$. This is the notation which was used by this author in Ref. [21]. Except for the inclusion of the fin effectiveness factor η_f^{-1} in θ_{conv}, Eq. (2.11) is the same as Eq. (2.1), which was derived for the case of infinitely conductive substrate (fin) material. In situations where the heat transfer from the prime surface cannot be neglected, it is customary [17] to define an overall fin effectiveness parameter $\eta = 1 - (A_{fin}/A_{total})(1 - \eta_f)$, but this refinement need not concern us here; for our high aspect ratios we approximate $\eta = \eta_f$.

R. W. Keyes [36] has recently proposed another approximate solution to Eqs. (2.2)-(2.8). Whereas we solve all but Eq. (2.6) exactly and satisfy Eq. (2.6) only in the integral, Keyes solves all but Eq. (2.7) exactly and satisfies Eq. (2.7) in the integral. Keyes' approach appears to be a better approximation when the fin efficiency is very low (i.e., $z \gg \lambda$) because it gives $T_w(x,z) \rightarrow T_A$ as $H \rightarrow \infty$, which is consistent with physical intuition. Specifically Keyes' solutions are:

$$T_w(x,z) = \dot{q}'' \cdot \frac{(w_c + w_w)\lambda^* \cdot \cosh((H-z)/\lambda^*)}{k_w w_w \cdot \sinh(H/\lambda^*)} \qquad (2.14)$$

$$\theta_c(x,z) = T_w(x,z) \cdot [x/(a+x)] ,$$

where $a = \rho C v w_c^2 / k_c Nu_x = D \cdot Re \cdot Pr / 4Nu_x$ and $\lambda^* = \lambda(1 + x/a)^{1/2}$.

Either set of approximations gives identical results for small H, as can be seen by expanding $\lambda \cdot \coth(H/\lambda)$ as a power series and retaining only the first two terms. For $H \leq \lambda$ the approximations differ by less than 0.5%. Since this was the region in which all our experiments were conducted (see Table 3-5), the two approximations were experimentally indistinguishable in this work. However, neither approximation represents the true physical situation, because the assumption that the energy conservation equations (2.6) or (2.7) are satisfied in the integral implies that some sort of thermal mixing in the fluid occurs in the z direction. Using a thermal-diffusion argument similar to that in Section 2.0.1, it can be shown that the silicon fins will indeed allow the coolant temperature profile to partially equilibrate in the z-direction, but <u>not</u> completely. Thus we expect that the use of Eq. (2.9) may lead to overly optimistic predictions. The exact behavior could be found by numerically solving Eqs. (2.2)-(2.8) on a computer, but this would provide little insight. Furthermore, simple analytical formulas are very desirable in the subsequent design work. Our approach in this work will be to modify slightly the formulas for thermal resistance used in our earlier work [21] so as to insure that the design is <u>conservative</u>, rather than trying to be analytically precise. We do this by approximating

$$\theta_{cal} = 1/\rho C f \eta . \qquad (2.15)$$

η is the fin effectiveness defined in Eq. (2.13):

$\eta = (\tanh m)/m$, where

$$m = (2h/k_w w_w)^{1/2} H = (Nu_L k_c/k_w)^{1/2}[(w_c + w_w)/2(w_c w_w)^{1/2}]\alpha \tag{2.16}$$

η is thus a monotonically decreasing function of m, with $\eta \simeq 1$ for $m \ll 1$, and $\eta \simeq m^{-1}$ for $m \gg 1$. Eq. (2.15) can be proved to be conservative; it was derived from the approximations that

$$T_w(x,z) = \dot{q}'' \cdot \frac{(w_c + w_w)\lambda \cdot \cosh((H-z)/\lambda)}{k_w w_w \cdot \sinh(H/\lambda)} (1 + x/a) \tag{2.17}$$

$$\Theta_c(x,z) = T_w(x,z) \cdot [x/(a+x)] . \tag{2.18}$$

Here all the equations (2.2)-(2.8) are satisfied except for Eq. (2.6), which can only be satisfied if h is assumed to vary along the fin length, always having a value less than that expected from the Nusselt number Nu_L. Since in fact the heat transfer is better than that, Eq. (2.15) must be conservative.

The three approximate theories for thermal resistance can be easily compared by using the dimensionless axial distance $L^* \equiv L/(D \cdot Re \cdot Pr)$. The peak thermal resistance (at $x = L$) is written

$$\theta = \Upsilon D/k_{eff} LW, \text{ where } k_{eff} = 2\sqrt{k_c Nu_L k_w w_c w_w/(w_c + w_w)}.$$

Here Υ is a dimensionless thermal resistance defined as follows:

For the "optimistic" theory used in Eq. (2.9) and in Ref. [21],

$$\Upsilon = \coth(H/\lambda) + 4Nu_L L^* \cdot (\lambda/H) . \tag{2.19}$$

For Keyes' theory (Eq. (2.14)),

$$\Upsilon = (1 + 4Nu_L L^*)^{1/2} \cdot \coth(H/[\lambda(1 + 4Nu_L L^*)^{1/2}]) . \tag{2.20}$$

For the "conservative" theory (Eq. (2.17)), **which is used in all subsequent calculations,**

$$\Upsilon = (1 + 4Nu_L L^*) \cdot \coth(H/\lambda) . \tag{2.21}$$

The theories all converge for the case where $4Nu_L L^* \ll 1$, i.e., when caloric thermal resistance is negligible.

The precise value of Nu_L depends on the channel geometry and how fully developed the thermal boundary layer is at the channel exit ($x = L$). We tentatively assume that, owing to our very narrow channel size, the flow will be laminar ($Re \lesssim 2100$); then from the discussion in Section 2.0.1 we have the following general asymptotic forms for Nu_L:

$Nu_L \propto (L^*)^{-1/3}$ for $L^* \ll 0.02$;

$Nu_L \simeq Nu_\infty$, a constant, for $L^* \gg 0.02$.

In the former case the thermal boundary layer is developing; in the latter case it is fully developed. Fig. 2-3 shows a plot of Nu_L vs. $L^* = L/(D \cdot Re \cdot Pr)$ for the case of constant heat flux into parallel planes (calculated from the published eigenvalues of Cess and Shaffer [37]).

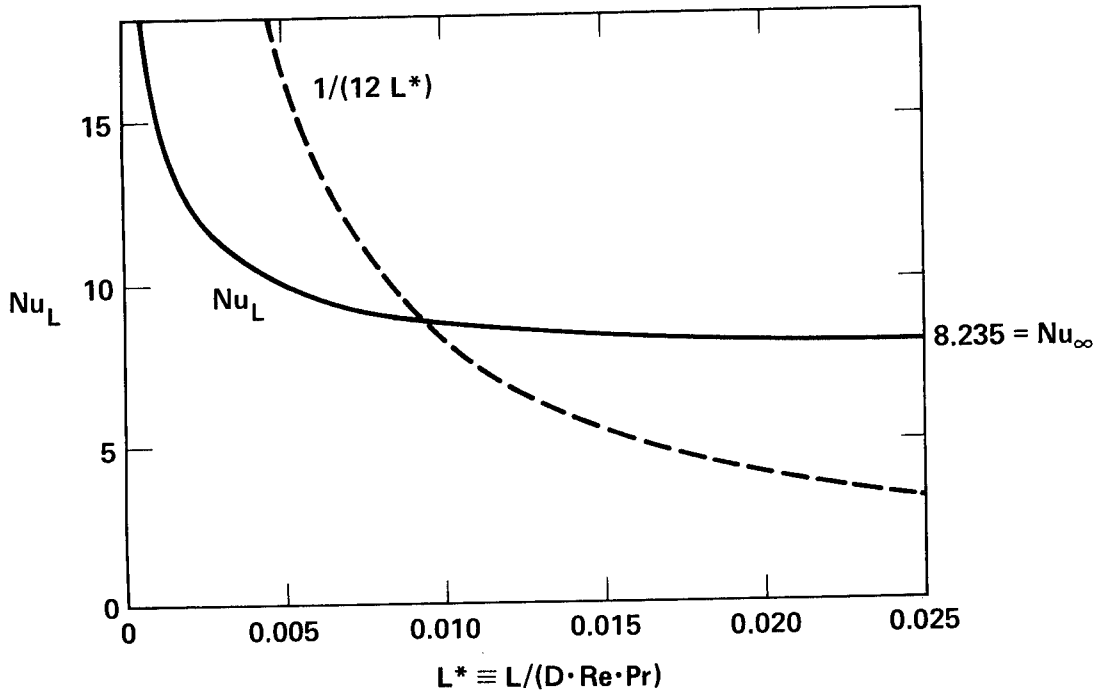


Figure 2-3: Local Nusselt number for laminar flow between parallel plates with uniform heat flux, as a function of dimensionless length $L^* \equiv L/(D \cdot Re \cdot Pr)$.

Not knowing *a priori* which region we are in, we conservatively assume that Nu has the minimum, asymptotic (large L) value Nu_∞; in any case the dependence of Nu on L is weak. As shown in Fig. 2-4, the exact value of Nu_∞ depends on the shape of the channel cross section, and for rectangular channels with all walls transferring heat (Case 1) ranges from 3.61 to 8.235. Case 2 is actually the one of interest, because the ends of the fins are capped by a thermally insulating cover plate.

In subsequent calculations we will frequently find the thermal resistance has the form

$$\theta = \theta_{conv} + \theta_{cal} = ax^{\alpha} + bx^{-\beta},$$

where a, b, α, and β are nonnegative constants, and x is a linear dimension such as hydraulic diameter. Minimizing θ would then be accomplished when

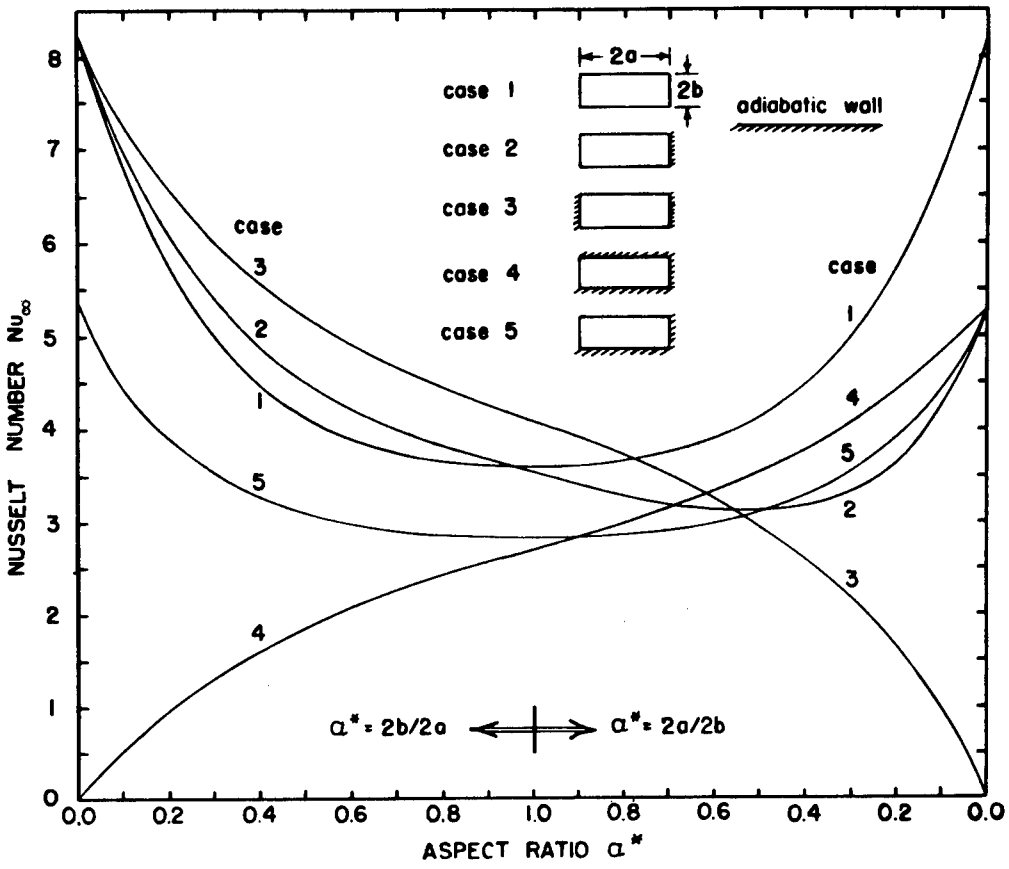


Figure 2-4: Uniform-flux Nusselt number for fully-developed laminar flow in rectangular ducts, with one or more walls transferring heat; Nu is based on wetted perimeter (figure courtesy of A. L. London [34]).

$$x = (\beta b/\alpha a)^{1/(\alpha+\beta)}, \tag{2.22}$$

for which

$$\theta_{min} = (1 + \alpha/\beta)\cdot ax^{\alpha} = (\alpha+\beta)[b^{\alpha}a^{\beta}/\alpha^{\alpha}\beta^{\beta}]^{1/(\alpha+\beta)}. \tag{2.23}$$

Note: the downstream Nusselt number Nu_L will henceforth be denoted as Nu, i.e., we drop the subscript L for convenience.

2.1.2.1. Constant-Pressure Constraint

As discussed, pump pressure or pump power constraints prevent us from making the channels arbitrarily narrow. We consider first the case where the pressure P is given. The pressure drop in a channel of hydraulic diameter D may be written as $P = (2L/D)(\rho v^2)c_{fm}$, where v is the mean flow velocity and c_{fm} is the mean friction factor, which will depend on the Reynolds number and the channel geometry. For a **fully-developed** laminar momentum boundary layer, $(c_{fm}\cdot Re)$ is a constant determined by channel geometry, as shown in Fig. 2-5;

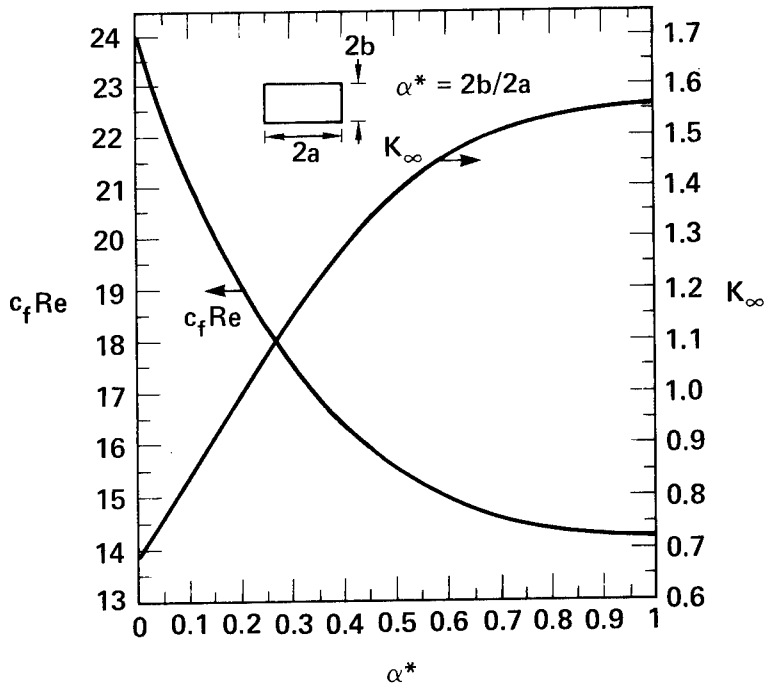


Figure 2-5: Normalized friction factor $\Phi \equiv c_f Re$ and entrance-effect loss factor K_∞ for fully-developed laminar flow (Ref. [34]).

we shall denote it by the letter Φ. Thus we have $\Phi \equiv (c_{fm} Re) = (c_{fm} vD\rho/\mu)$, whence $P = 2\Phi\mu Lv/D^2$. The total volume flow rate in our rectangular channels is easily seen to be

$$f = vWD\alpha/4 = WPD^3\alpha/8\Phi\mu L\,, \tag{2.24}$$

whence, from Eq. (2.15),

$$\theta_{cal} = 1/\rho Cf\eta = (8\Phi\mu L/\rho CPW)\cdot(D^{-3}\alpha^{-1}\eta^{-1})\,. \tag{2.25}$$

The convective thermal resistance is, from Eq. (2.12) or (2.21),

$$\theta_{conv} = (1/k_c NuLW)\cdot(D\alpha^{-1}\eta^{-1}) \tag{2.26}$$

It is convenient to treat (D, α, and w_w) as independent variables; obviously they are related in a one-to-one correspondence to the geometrical parameters w_c, w_w, and H. We seek an optimum design, i.e., one which minimizes the maximum thermal resistance $\theta = \theta_{conv}(w_w,D,\alpha) + \theta_{cal}(w_w,D,\alpha)$. This is achieved by solving $\partial\theta/\partial\alpha = \partial\theta/\partial D = \partial\theta/\partial w_w = 0$. As a first step, we solve for $\partial\theta/\partial w_w = 0$; by inspection of Eqs. (2.16), (2.25), and (2.26), this occurs when η is maximized, i.e., when $\partial m/\partial w_w = 0$. This can only be satisfied if $w_w = w_c$ or if $\partial w_c/\partial w_w = w_c/w_w$. However, the latter possibility leads to the unphysical relation $w_c + w_w = 0$. Thus we conclude that $w_w = w_c$.

Both θ_{conv} and θ_{cal} decrease monotonically with increasing α, so the theoretical optimum

value for α is infinite. The fin efficiency η rolls off as α^{-1} for large α, hence as $\alpha \to \infty$, $(\alpha\eta)$ asymptotically approaches an upper limit which we denote

$$\alpha_c \equiv [k_w/(k_c Nu)]^{1/2}. \tag{2.27}$$

α_c can be viewed as the maximum enhancement in effective heat-transfer surface which may be achieved using fins. For very high aspect ratios, we have Nu = 140/17 = 8.235, hence α_c = 5.43 for a water-cooled silicon substrate.

For $w_c = w_w$ and $\alpha = \infty$, we now have

$$\theta_{conv} = (1/k_c NuLW\alpha_c)D\,, \tag{2.28}$$

$$\theta_{cal} = (8\mu\Phi L/\rho CPW\alpha_c)D^{-3}. \tag{2.29}$$

Setting $\partial\theta/\partial D = 0$ yields an optimum channel dimension D which minimizes θ (refer to Eq. (2.22)):

$$D = (24\mu\Phi k_c NuL^2/\rho CP)^{1/4}\,; \tag{2.30}$$

θ_{conv} and θ_{cal} are in a 3:1 ratio for this value of D. From Eq. (2.23), we have the optimized thermal resistance

$$\theta_{opt} = (4/3)\theta_{conv} = 2.95(\mu\Phi/k_c Nuk_w^2 L^2 W^4 \rho CP)^{1/4}\,. \tag{2.31}$$

Using our optimized value for D, we calculate that

$$L/(D \cdot Re \cdot Pr) = 1/(12Nu) = 0.010\,,$$

which implies an **almost** fully-developed temperature profile. A more precise estimate of Nu may be obtained by correcting for the thermal entrance effects. Solving the relation $L/(D \cdot Re \cdot Pr) = 1/(12Nu)$ graphically, as shown in Fig. 2-3, we conclude that $Nu_L = 8.882$, which is only 8% larger than our original estimate. For a room-temperature water-cooled silicon heat sink on a (1 cm) $\times$ (1 cm) substrate, with a water pressure of P = 50 psi = 3.45×10^6 dynes/cm^2, our design procedure gives D = 116.6 μm, hence the conservatively designed optimum is:

$w_c = w_w = D/2 = 58\ \mu m$;

$H = \infty$;

$\theta_{opt} = 0.055^{\circ}C/W$.

Note that $Re = 5.42(L^2\rho k_c^3 Nu^3 P/\mu^5 \Phi C^3)^{1/4} \simeq 1560$ under the stated conditions, hence the flow will indeed be laminar.

If we use Keyes' approximations (Eq. (2.20)), we would operate at $L/(D \cdot Re \cdot Pr) = 1/(4Nu)$, where $Nu \simeq 8.235$. D is calculated to be smaller by a factor of 0.75, i.e., a 44-μm channel and wall width. His predicted thermal resistance is 17% smaller. This difference is expected because our design formulas were chosen to be conservative (pessimistic) in their calculation of θ.

2.1.2.2. Constant-Pressure, Constant-Fin-Height Constraint

In the preceding paragraphs, we approximately optimized the heat sink geometry by assuming a constant coolant supply pressure P. The conclusion that the aspect ratio should be infinite implies an infinite flow rate and infinite pumping power, which is unphysical. Clearly one could truncate the fins (i.e., fix H to be some finite value) and optimize for this added constraint, in which case a finite-flow condition would result. If this truncation is performed at a point where the fin efficiency would be low, (i.e., an aspect ratio much higher than α_c), then very little heat-sinking performance penalty would be incurred.

To analyze this, let H be fixed; then the remaining independent variables are w_c and w_w. For this section, we approximate $D \simeq 2w_c$ and $\alpha = 2H/(w_c + w_w)$. Setting $\partial\theta/\partial w_w = 0$ implies that $\partial(\alpha\eta)/\partial w_w = 0$, which leads to the conclusion

$$\frac{w_c - w_w}{w_c + w_w} = \frac{N}{(\sinh m)(\cosh m)} = 2m \cdot \operatorname{csch} 2m \qquad (2.32)$$

where $m = H\sqrt{Nu_L k_c / k_w w_c w_w}$, as defined in Eq. (2.16). (A similar equation was obtained by Keyes using his set of approximations.) Eq. (2.32) implies that $w_c < w_w$ for finite values of H. If Eq. (2.32) is satisfied, then $\partial(\alpha\eta)/\partial w_c = 0$ because $(\alpha\eta)$ is symmetric with respect to interchange of w_c and w_w. Thus setting $\partial\theta/\partial w_c = 0$ leads to

$$2w_c = (24\mu\Phi k_c NuL^2/\rho CP)^{1/4}, \qquad (2.33)$$

which is identical to Eq. (2.30). Thus for a truncated (finite-H) heat sink, the calculated optimum channel width w_c is the same as for the infinite-H case, but the fin width w_w is smaller in accordance with Eq. (2.32). The optimized thermal resistance θ_H is increased by a factor

$$\theta_H/\theta_\infty = (w_c/w_w)^{1/2} \cdot [(w_c + w_w)/2w_c] \coth m,$$

where θ_∞ is for the case of infinite fin height, calculated in Eq. (2.31). Table 2-1 summarizes this result for 3 different values of m: Fig. 2-6 is a plot of optimized thermal resistance as a function of H, normalized to the case $H = \infty$. We see that $H \simeq 1.32\alpha_c w_c$ is a good place to

m	w_w/w_c	θ_H/θ_∞	$H/w_c\alpha_c$
1	0.2891	1.574	0.538
1.5	0.5391	1.1579	1.1013
2	0.7443	1.0486	1.7255

Table 2-1: Optimized dimensions and thermal resistance for various fixed fin heights H.

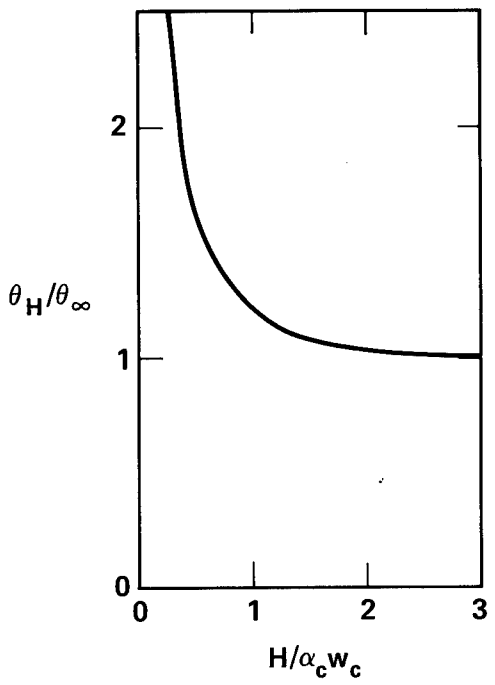


Figure 2-6: Optimized thermal resistance as a function of fin height.

truncate the fins, because the thermal resistance is only 10% above its asymptotic limit. For our 50-psi inlet water pressure, $1.32\alpha_c w_c \simeq (1.32)\cdot(5.43)\cdot(58\ \mu m) = 416\ \mu m$, which conveniently is less than a typical silicon wafer thickness!

2.1.2.3. Constant-Pumping-Power Constraint

The preceding discussion required a somewhat arbitrary truncation of the fins at some fixed height, in order to arrive at a physically realizable solution. Clearly one does not wish the fin aspect ratio to be very much larger than α_c, because the fluid flowing past the tops of the fins is wasted (we are expending mechanical energy to pump it, but it is transferring almost none of the heat). Thus an alternative, more fundamental boundary condition would

be to maximize the heat transferred by a given amount of mechanical work. That is, we shall optimize for a constant pumping power $\dot{Q} = Pf$. Recalling from Eq. (2.24) that $f = WPD^3\alpha/8\Phi\mu L$, we find $f^2 = W\dot{Q}D^3\alpha/8\Phi\mu L$, so we can recast Eq. (2.25) in terms of $\dot{Q}$:

$$\theta_{cal} = (8\Phi\mu L/W\dot{Q}\rho^2C^2)^{1/2} \cdot D^{-3/2}\alpha^{-1/2}\eta^{-1}. \tag{2.34}$$

Referring to Eqs. (2.16), (2.28) and (2.34), the total thermal resistance $\theta = \theta_{cal} + \theta_{conv}$ is a function of D, α, and w_w; setting $\partial\theta/\partial w_w = 0$ again leads to the conclusion that $w_w = w_c$. Substituting in this result and setting $\partial\theta/\partial D = 0$ then leads to the relationship (from Eq. (2.22)) $D = (18\mu\Phi Nu^2k_c^{\,2}L^3W/\rho^2C^2\dot{Q})^{1/5} \cdot \alpha^{1/5}$. Substituting this in, we find that $\theta \propto \alpha^{1/5}\coth(\alpha/\alpha_c)$, where α_c was defined in Eq. (2.27). Setting $\partial\theta/\partial\alpha = 0$ leads to the relation $(\sinh \alpha/\alpha_c) \cdot (\cosh \alpha/\alpha_c) = 5\alpha/\alpha_c$, which when solved numerically yields

$$\alpha = 1.789\alpha_c = 1.789[k_w/(k_cNu)]^{1/2}. \tag{2.35}$$

The fin effectiveness is therefore $\eta = 0.529$. The optimum channel hydraulic diameter is

$$D = 2.00(\mu\Phi k_c^{\,1.5}Nu^{1.5}k_w^{\,0.5}L^3W/\rho^2C^2\dot{Q})^{1/5}; \tag{2.36}$$

θ_{conv} and θ_{cal} are in a 3:2 ratio for this value of D. Thus the optimized thermal resistance is

$$\theta_{opt} = (5/3)\theta_{conv} = 3.53(\mu\Phi/NuL^2W^4k_w^{\,2}k_c\rho^2C^2\dot{Q})^{1/5}. \tag{2.37}$$

In this design we find that $L/(D \cdot Re \cdot Pr) = 1/(6Nu) = 0.025$; the temperature profile is thus even more nearly fully-developed than in the constant-pressure case and no entry-length correction to Nu is necessary. For water-cooled silicon heat sinks, we have used the data in Fig. 2-4 to solve Eq. (2.35); the solution is $\alpha = 10.6$, for which the channel aspect ratio is $H/w_c = 1/\alpha^* = \alpha - 1 = 9.6$. It is interesting that this aspect ratio is independent of the hydraulic power $\dot{Q}$; it is determined only by the ratio $k_w/(k_cNu)$. Using $\Phi = 21.0$ (from Fig. 2-5), we conclude that for a pumping power of $\dot{Q} = 6.22$ Watts, we have $D = 88.9\ \mu m$, so the optimized design is

$w_c = w_w = D\alpha/2(\alpha - 1) = 49.1\ \mu m$;

$H = (\alpha - 1)w_c = 471\ \mu m$;

$\theta_{opt} = 0.063^\circ C/W$ at $f = 18.0\ cm^3/sec$, $P = 3.45 \times 10^6$ dynes/cm^2 (50 psi).

Here $Re = 2.991(L^2\rho^2k_c^{\,3.5}Nu^{3.5}\dot{Q}/\mu^6\Phi WC^3k_w^{\,0.5})^{1/5} = 900$, which indeed corresponds to laminar flow. We see that for the same pressure as in the previous analyses, a 15% performance penalty (increase in θ_{opt}) was paid for optimizing for a **finite** (as opposed to

infinite) pump power $\dot{Q}$. In fact it can be shown in general that the penalty is a factor $1.113(\Phi_{\dot{Q}}Nu_P/\Phi_P Nu_{\dot{Q}})^{1/4}$, where the subscripts $\dot{Q}$ and P refer respectively to the values obtained in the constant-pumping-power and constant-pressure cases.

It is instructive to rewrite Eq. (2.37) in terms of power fluxes (power per unit area; W/cm^2). Specifically, we define the normalized thermal resistance $R_{opt} = LW\theta_{opt}$; this is the thermal resistance for a unit of substrate area. We further define the normalized mechanical pumping power $\dot{Q}'' = \dot{Q}/LW$; this is the amount of mechanical energy being expended to cool a unit substrate area. Then we have

$$R = 3.53(\mu L^2/k_c k_w^2 \rho^2 C^2 \dot{Q}'')^{1/5} \quad . \tag{2.38}$$

Interestingly, for a given, finite value of pumping power flux $\dot{Q}''$, the normalized thermal resistance can be reduced without limit by reducing the manifold spacing L. We shall show in Section 2.2.3 that header losses (manifold pressure drops) are not dominant in our designs, regardless of channel length. Thus there is no immediately obvious theoretical bound on the heat transfer rates which may be achieved from a planar surface using optimized laminar-flow heat exchangers, provided one is willing to make the channel length very short. This idea is discussed in more detail in Section 2.2.6.

2.1.3. Discussion

We see that typical channel widths of about 50 μm are calculated using our first-order optimization procedure, and predict that such microscopic laminar-flow heat exchangers would provide excellent heat transfer. Because the thermal resistances are substantially less than $0.1^{\circ}C/W$, power fluxes of more than 1000 W/cm^2 should be readily achievable while maintaining surface temperatures of less than $100^{\circ}C$. It is instructive to pause here to examine Eqs. (2.31) and (2.37) to see the effects of various parameters on θ_{opt}, and to determine where possible first-order improvements in heat transfer might be achieved. Admittedly a number of approximations have been made in arriving at these formulas; on the other hand, they have a transparency and simplicity which a more accurate numerical analysis might obscure. Note that in either the constant-pressure or constant-pumping-power optimizations, most parameters affect the thermal resistance as a 1/4 or 1/5 power. This is a consequence of the fact that the thermal resistances due to convection, fin conduction, and coolant heating are all comparable in an optimized structure, so the sensitivity of the total thermal resistance to a single parameter is reduced. Note also that a very large expenditure of effort (32 times the pumping power or 16 times the pressure) would be required to improve

the heat transfer by a factor of 2; conversely, the penalty paid for a substantial reduction in pressure (say, down to 2 psi) or pumping power is not large.

It is useful to factor out from θ_{opt}^{-1} the parameters which pertain to the coolant fluid; in this way, we can define a "coolant figure of merit" (CFOM). In the constant-pressure case, $CFOM_P = (k_c \rho C/\mu)^{1/4}$; in the constant-pumping-power case, $CFOM_{\dot{Q}} = (k_c \rho^2 C^2/\mu)^{1/5}$. Table 2-2 tabulates the CFOMs (relative to water) for a variety of potential coolants.

Fluid	k_c (W/cm-K)	ρC (J/cm^3-K)	μ (centipoise)	$CFOM_P$	$CFOM_{\dot{Q}}$
Air	.00026	.00118	0.0184	0.154	0.044
Air @ 100 atm	.00026	.118	0.0184	0.487	0.110
Helium gas	.00145	.00086	0.0194	0.215	0.054
Water	.00609	4.164	0.8513	1.000	1.000
Fluorinert® FC-77	.00064	1.862	1.42	0.409	0.417
Mercury	.0830	1.880	1.527	1.361	1.091
Liquid N_2 @77°K	.00140	1.62	0.165	0.824	0.709

Table 2-2: Coolant Figure of Merit (CFOM) for several fluids at 20-27°C and 1 atm (unless noted), normalized so that CFOM = 1 for water. The subscripts P and $\dot{Q}$ denote constant-pressure and constant-pumping-power optimization, respectively.

Water has the best performance of the practical coolants; mercury is only slightly better but much less practical. The Fluorinert® family of coolants are very inert and hence might be preferred in cooling applications where water cannot be used due to reliability or corrosion problems. However, the optimized heat transfer in this latter case is only about 40% of that achieved with water. Air and helium have also been included as possible coolants, mainly to show their inferior performance (typically one-tenth the CFOM of water) at atmospheric pressure. Note that this analysis was based on **incompressible** fluid mechanics and so would only be accurate for small relative pressure drops ($P_{in} - P_{out} \ll P_{out}$) and low channel velocities (Mach number M<0.6). Their performance is better at elevated pressures (say, 100 atm), owing to their higher density and hence higher volumetric heat capacity ρC. On the other hand, such high pressures require more robust packaging; even though the pressure **drop** might be only 2 atm (as in our water-cooled heat sink design), one must nonetheless design a 100-atm package, which would not be necessary in the liquid-cooled case.

The substrate thermal conductivity k_w enters into the thermal resistance as $k_w^{-1/2}$ for the constant-pressure optimization, or $k_w^{-2/5}$ for constant pumping power. Referring back to Table 1-2, we conclude that the use of copper affords approximately a 64% improvement in heat transfer compared with silicon. Using a GaAs substrate results in approximately a 40% **reduction** in heat transfer. It is interesting that the thermal conductivity of silicon is about 14.5 W/cm-K at 77°K, ten times larger than at room temperature (Fig. 1-2). Considering that the CFOM of liquid nitrogen is only slightly smaller than for water, very low thermal resistances could be achieved at cryogenic temperatures by cooling silicon substrates with liquid nitrogen.

The channel length L (the manifold separation) enters into the optimized thermal resistance θ_{opt} as $L^{-1/2}$ or $L^{-2/5}$. The normalized thermal resistance $R_{opt} = LW\theta_{opt}$ thus decreases with decreasing channel length as $L^{1/2}$ or $L^{3/5}$. While we have been assuming that the substrate length is 1 cm, there is no requirement that the channels have to extend the entire substrate length. One could, for example, construct 10 consecutive sets of channels, each set 1 mm in length, within a 1-cm long substrate. Of course manifolds (headers) must be fashioned at the ends of each set of channels. Fig. 2-10 (on page 47) shows a possible scheme for doing this. By optimizing the channels for L = 1 mm rather than 10 mm, we could more than triple the heat transfer from the substrate.

The final term in Eqs. (2.31) and (2.37) to be discussed is the Nusselt number, which appears as $Nu^{-1/4}$ or $Nu^{-1/5}$ in the thermal resistance formulas. Recall that we assumed that Nu had its lowest possible value (fully-developed thermal boundary layer), in accordance with the data in Fig. 2-4. We shall examine this assumption in more detail in the next section, but for now we simply note that <u>if</u> Nu could be increased <u>without</u> increasing the friction coefficient c_f, then a factor of $Nu^{1/4}$ or $Nu^{1/5}$ decrease in thermal resistance would accrue, which is not a very impressive gain. In fact, however, most approaches which increase Nu would also increase c_f. This obviously adversely affects the performance which could be achieved, so $Nu^{1/4}$ or $Nu^{1/5}$ is clearly an upper bound on the heat-transfer improvements which may be achieved, within the constructs of our model. However, our design optimization procedure breaks down when $Nu \geq k_w/k_c \simeq 250$ for water-cooled silicon, or 660 for water-cooled copper. This is because it weights both the prime and extended (finned) surfaces by a fin efficiency factor η, whereas in fact the heat transfer from the prime surface is 100% effective. This is not a serious approximation when the Nusselt number is relatively low. Section 2.2.2 deals with the case where the Nusselt number is very large, and hence the prime surface is the dominant heat-transfer surface.

This upper bound may be tightened further if we have some idea of the amount by which the friction factor will increase. Suppose we have a real heat sink which, for a given pumping power $\dot{Q}$, has already been optimized to provide the minimum possible peak surface temperature; we denote this minimum value as T_o. This optimum design will be operating at some Reynolds number Re_o, and exhibit some friction coefficient c_{fo} and Nusselt number Nu_o at that Reynolds number. Now our optimization procedure assumed that both Nu and $\Phi \equiv c_f Re$ are constants; in the real heat sink, the functional form would not be that simple. Substituting in Nu_o and $\Phi_o \equiv c_{fo} Re_o$ into our optimization, we would calculate a thermal resistance which is at least as low as in the "real" heat sink. This is because the operating point of the real heat sink is a self-consistent solution of Eqs. (2.16), (2.28) and (2.34); applying the optimization procedure of Section 2.1.2.3 would yield either the same operating point, or one with even lower θ. Referring to Eq. (2.37), we therefore have a lower bound on the thermal performance of our real heat sink:

$$\theta_{real} \geq 3.53(\mu c_{fo} Re_o / k_c Nu_o k_w^2 L^2 W^4 \rho^2 C^2 \dot{Q})^{1/5} \quad .$$

Thus the percentage performance improvement over the laminar-flow, fully-developed temperature profile case would be bounded by the percentage improvement in $(Nu/c_f Re)^{1/5}$ over the laminar case (for which it has the approximate value $(6.7/21.3)^{1/5} = 0.79$). Now consider the Colburn analogy [38], which states that

$$j_H \simeq c_f/2 \quad \text{where } j_H \equiv Nu/(RePr^{1/3}).$$

Thus $(Nu/c_f Re)^{1/5} \simeq 0.87 Pr^{1/15} \simeq 0.94$ for warm (55°C) water, which is only a slight (19%) improvement over the laminar-flow value of 0.79. The Colburn analogy holds very well for highly turbulent flow in smooth pipes. When it does fail, it is usually due to form drag and hence a **higher-than-normal** friction factor would result, which will further degrade performance. Only in unusual circumstances (e.g., certain spirally fluted tubes [39], or the use of surface spoilers [40]) does it fail in the other (high-Nu) direction, and then only by a factor of 30% or so, which would result in a 5% performance improvement. Thus turbulent flow does not appear to provide a significant advantage when compared with our optimized laminar-flow heat exchangers, provided $Nu \lesssim k_w/k_c$. (If $Nu \gtrsim k_w/k_c$, the prime surface becomes the dominant heat-transfer surface; this is discussed further in Section 2.2.2.)

As confirmation of this assertion, the reader is referred to "Compact Heat Exchangers" by Kays and London [17], in which experimental values of j_H and c_f are plotted against Reynolds number for a very large number of different heat-exchanger structures, many of which contained highly complex structures to interrupt the thermal boundary layer. In all cases one finds that $c_f/2$ exceeds j_H; in many examples $c_f/2$ exceeds j_H by a factor of 2 or more.

In the absence of other constraints such as manufacturing limitations, the superiority of laminar-flow heat exchangers (especially parallel-plate geometries) from an energy-expenditure viewpoint has long been recognized in the heat-transfer community [34, 41], but manufacturing constraints and tolerances have traditionally prevented the extreme miniaturization which our design calls for. In Section 3.1.1 we shall describe how silicon micromachining techniques make implementation of our design entirely feasible.

2.2. Refinements

In this section we examine some of the assumptions and approximations which were explicitly or implicitly made in our first-order design for the constant-pumping-power constraint. The purpose here is not to construct a comprehensive theory which accounts for all 2nd-order effects and makes no approximations. Rather, our intent is to confirm that these assumptions and approximations were reasonably good, and to estimate the errors which are incurred in our analysis as a result of these assumptions. We will need to know the magnitude of these errors when we compare our experimental results with theory. Particular attention has been paid to insuring that the self-consistency of our assumptions does not result in circular arguments when optimizing the heat sink structures; see the discussions of developing thermal boundary layer and of turbulent flow. The discussions will be limited to the analysis of considerations which arise in the design of water-cooled silicon heat sinks.

2.2.1. Developing Thermal Boundary Layer

In the elementary constant-pumping-power optimization of Section 2.1.2.3, we **assumed** that the downstream ($x = L$) Nusselt number Nu_L was a constant (approximately equal to 6.7 for the 10:1 aspect ratio in silicon/water heat sinks), as is the case when the laminar thermal boundary layer is fully developed. It was then concluded that an optimum design would have $L/(D \cdot Re \cdot Pr) = 1/(6Nu_L) \simeq 0.025$, which does in fact imply a fully-developed temperature profile (Fig. 2-3). To ensure that we have not made a circular argument, we consider now the possibility that a better optimum exists in the regime of a developing laminar thermal boundary layer, i.e., $L/(D \cdot Re \cdot Pr \ll 0.01$. We shall show that the best performance is achieved at the fully-developed end of this regime.

The asymptotic form of Nu_L for rectangular tubes of length L is $Nu_L \simeq \beta[L/(D \cdot Re \cdot Pr)]^{-1/3}$ where $\beta \simeq 1.2$ to 1.5 for aspect ratios ranging from 1 to ∞ [34]. Now $Re = vD\rho/\mu$, where $v = (2D\dot{Q}/\mu\Phi LW\alpha)^{1/2}$; thus

$$Nu_L = \beta(2\rho^2C^2\dot{Q}/k_c^2L^3\mu\Phi W)^{1/6}D^{5/6}\alpha^{-1/6}. \qquad (2.39)$$

Substituting into Eq. (2.26) yields:

$$\theta_{conv} = AD^{1/6}\alpha^{-5/6}\eta^{-1} = AD^{1/6}\alpha^{1/6}(\alpha\eta)^{-1}, \text{ where } A \equiv \beta^{-1}(\mu\Phi/2k_c^4L^3W^5\rho^2C^2\dot{Q})^{1/6}.$$

From Eq. (2.34), we write

$$\theta_{cal} = BD^{-3/2}\alpha^{-1/2}\eta^{-1} = BD^{-3/2}\alpha^{1/2}(\alpha\eta)^{-1}, \text{ where } B \equiv (8\mu\Phi L/W\dot{Q}\rho^2C^2)^{1/2}.$$

We wish to minimize $\theta = \theta_{conv} + \theta_{cal}$, subject to the constant-pumping-power constraint. As before, setting $\partial\theta/\partial w_w = 0$ implies $w_c = w_w$. Then

$$(\alpha\eta) = \alpha(\tanh m)/m = \alpha_c \tanh(\alpha/\alpha_c),$$

but unlike before, $\alpha_c = (k_w/k_cNu_L)^{1/2}$ is not a constant, because Nu_L is not constant but depends on D and α in accordance with Eq. (2.39).

It is convenient to define a variable $\varepsilon = \alpha/\alpha_c$; the independent variables will now be D and ε, rather than D and α. Now $(\alpha\eta)^{-1} = \alpha_c^{-1}\coth\ \varepsilon = ED^{5/12}\alpha^{-1/12}\coth\ \varepsilon$, where $E \equiv (kc\beta/k_w)^{1/2}(2\rho^2C^2\dot{Q}/k_c^2L^3W\mu\Phi)^{1/12}$. Furthermore $\alpha = E^{-12/11}D^{-5/11}\varepsilon^{12/11}$. Thus we have

$$\theta = \theta_{conv} + \theta_{cal} = AD^{7/12}\alpha^{1/12}E\coth\varepsilon + BD^{-13/2}\alpha^{5/12}E\coth\varepsilon =$$

$$= (AE^{10/11})D^{6/11}\varepsilon^{1/11}\coth\varepsilon + (BE^{6/11})D^{-14/11}\varepsilon^{5/11}\coth\varepsilon.$$

Setting $\partial\theta/\partial D = 0$ leads to the relationship $D = (7B/3A)^{11/20}(\varepsilon/E)^{1/5}$. Substituting this in, we find that $\theta \propto \varepsilon^{1/5}\coth\varepsilon$; setting $\partial\theta/\partial\varepsilon = 0$ yields $\varepsilon = 1.789$. At this design point,

$$L/(D\cdot Re\cdot Pr) = 3/(28Nu_L) = [3/(28\beta)]^{3/2} \simeq 0.027 \text{ to } 0.019\ ;\ Nu_L = 4.0 \text{ to } 5.6.$$

But this is almost exactly the same design point as we obtained by assuming Nu_L to be a constant (the fully-developed region). Thus, <u>either</u> asymptotic expression for Nu_L on page 21 leads to the same conclusion: the optimum design point of a laminar-flow heat sink occurs at around $L/(D\cdot Re\cdot Pr) \simeq 0.02$, which corresponds to an almost fully-developed temperature profile.

Although we assumed the hydrodynamic boundary layer was fully developed ($\Phi = c_fRe$ = constant), this is only true provided that $L/(D\cdot Re) > 0.05$. This is the case in our design provided $Pr \gg 2.5$, which is true for nonmetallic liquids. For $L/(D\cdot Re) \ll 0.05$, we would have $c_fRe \propto [L/(D\cdot Re)]^{-1/2}$. This would obviously make the thermal performance even worse for a given fixed power expenditure $\dot{Q}$. But we have just shown that the optimum performance obtainable by assuming a **developing** thermal boundary layer is essentially

identical to that obtained by assuming a **fully-developed** thermal boundary layer. Thus there is no possibility that incorporating into our analysis the extra flow-friction associated with the hydrodynamic entry length would lead to a better design than our fully-developed laminar-flow design.

The preceding argument has been repeated for the constant-pressure constraint and leads to the same conclusion, namely that designing in the developing thermal boundary layer regime cannot result in better performance than was calculated in Section 2.1.2.1. The exact result was that $L/(D \cdot Re \cdot Pr) = 0.00675 \cdot \Phi^{1/2}(Pr/\beta)^{3/2}$ which, for $\Phi = 24$, $Pr = 5$, and $\beta = 1.5$, gives $L/(D \cdot Re \cdot Pr) = 0.20$, which is well into the fully-developed regime. Since the asymptotic expression used for the developing thermal boundary layer was only valid in the region $(L/D \cdot Re \cdot Pr) \lesssim 0.01$, then the best design in this region will occur at the endpoint of this region, i.e., at $L/(D \cdot Re \cdot Pr) \simeq 0.01$, which is what we had concluded anyway based on the assumption that Nu was constant.

This result that the thermal boundary is nearly fully-developed is not an accident, but is in fact a direct consequence of the optimization of the sum $\theta = \theta_{conv} + \theta_{cal}$. The optimum design occurs when θ_{cal} is beginning to be comparable to θ_{conv} at the downstream end. That is, the temperature rise associated with coolant heating is comparable to the temperature drop in the channel cross section. It is intuitively clear that this condition would imply a fully-developed temperature profile. More mathematically, one can calculate that $\theta_{cal}/\theta_{conv} = 4NuL/(D \cdot Re \cdot Pr)$ for high-aspect-ratio structures, regardless of specific geometry. Thus having θ_{cal} comparable to θ_{conv} implies that $(L/D \cdot Re \cdot Pr) \simeq 1/4Nu$, which for laminar flow implies nearly-developed temperature profiles.

2.2.2. Low-Aspect-Ratio Designs (Turbulent Flow)

In Section 2.1.3, we calculated an upper bound on the performance increase (if any) which could be achieved (for a given pump power) by inducing turbulent flow in a microscopic finned heat sink. This bound was the factor $(Nu/c_f Re)^{1/5}/0.79 = 1.1(2j_H/c_f)^{1/5}Pr^{1/15}$, where $j_H = Nu/(RePr^{1/3})$ is the Colburn factor. The highest value of $2j_H/c_f$ which we have seen in the literature [40] is 1.3, corresponding to at most a 25% improvement. However, the validity of this result is dependent on the validity of our extended-surface heat-transfer model (in particular, the weighting of the heat transfer by a fin efficiency η). As discussed in Section 2.1.2, this model breaks down when the Nusselt number is high ($Nu > k_w/k_c \simeq 250$ for water-cooled Si; 660 for water-cooled Cu). In such cases, the heat transfer into the liquid is so

effective that essentially no heat would conduct up the fins. Thus we expect that the best design for the very high-Nusselt-number regime ($Nu \gg k_w/k_c$) would be a parallel-plate arrangement as shown in Fig. 2-7; no surface area enhancement is used because it would not be helpful and it would take up space which could be used for direct convective heat transfer into the fluid. Possibly one might include small turbulence promoters on the heated surface to get a higher Nusselt number at a lower Reynolds number, but these structures would not themselves transfer significant heat by conduction.

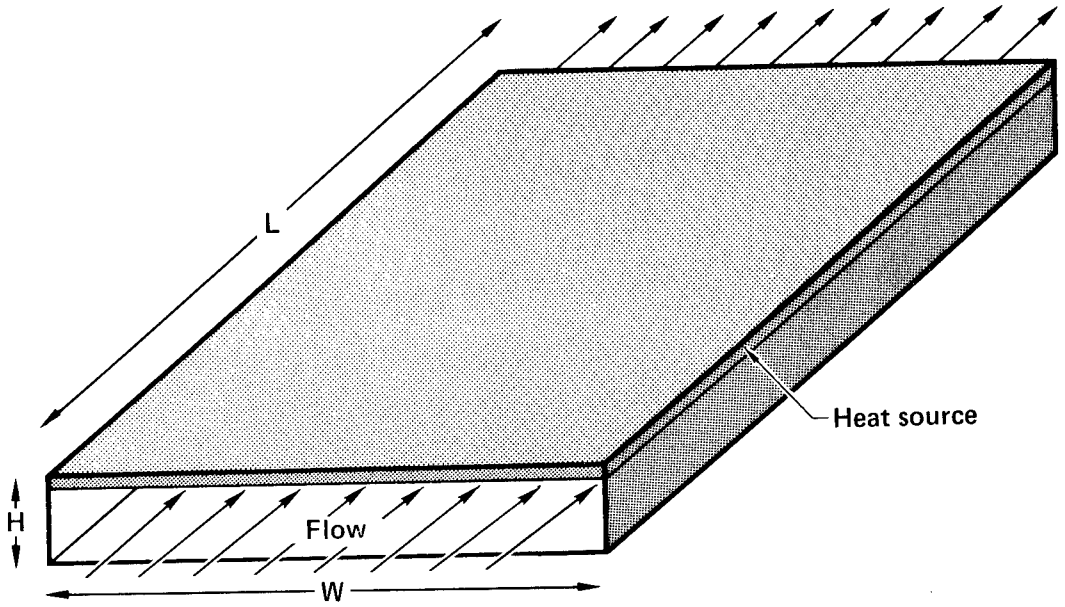


Figure 2-7: A simple turbulent-flow cooling duct would be optimal if $Nu \gg k_w/k_c$.

We shall analyze the structure in Fig. 2-7 by obtaining an **upper bound** on its potential performance for a given pumping power Q. Neglecting the effects of turbulence promoters on the amount of surface area, and assuming the gap H to be small relative to the plate width W, we have a hydraulic diameter of $D = 2H$. The pressure drop is then $P = (L/H)c_f\rho v^2$, where c_f is the friction factor. The volume flow rate is $f = vHw$, so the pumping power is $Q = Pf = LWc_f\rho v^3$, where $v = (Q/LW\rho c_f)^{1/3}$. The caloric thermal resistance is

$$\theta_{cal} = 1/\rho Cf = (L/\rho^2 C^3 W^2 Q)^{1/3} c_f^{\,1/3} H^{-1} \; . \; . \qquad (2.40)$$

It is convenient to express the convective heat transfer coefficient h in terms of the Colburn factor $j_H = Nu/(RePr^{1/3})$:

$$h = k_c Nu/D = k_c RePr^{1/3} j_H/D = k_c Pr^{1/3} j_H \rho v/\mu \; .$$

Substituting in for v, we find the convective thermal resistance:

$$\theta_{conv} = h^{-1} = (\mu^2 c_f / k_c^2 \rho^2 C j_H^3 L^2 W^2 Q)^{1/3} \quad . \tag{2.41}$$

Now it is an empirical fact that at the high Nusselt numbers which we are considering, the friction factor c_f and Colburn factor j_H are relatively insensitive functions of the Reynolds number (e.g., see Fig. 2-8). Thus an optimization procedure based on the assumption that c_f and j_H are constant would probably not be a bad approximation. Furthermore, suppose that we knew the **exact** functional forms of c_f and j_H as a function of Re; we could then in principle compute the optimal operating point, at which point $c_f = c_{fo}$ and $j_H = j_{Ho}$ are constants. If we now take these values and minimize $\theta_{turb} = \theta_{conv} + \theta_{cal}$ under the assumption that $c_f = c_{fo}$ and $j_H = j_{Ho}$ **independent** of Reynolds number, then we would end up with a calculated thermal resistance which is <u>at least</u> as good as the true situation, because the true operating point is a member of the space of possible solutions from which we are optimizing. Hence this approach would calculate an upper bound on the possible performance.

Referring to Eqs. (2.40) and (2.41), we find that for c_f and j_H constant, the optimum design is to let H be large; then $\theta \rightarrow \theta_{conv}$. Thus an upper bound on the performance (i.e., a lower bound on thermal resistance θ_{turb}) of our parallel-plate design is

$$\theta_{turb} > (\mu^2 / k_c^2 \rho^2 C L^2 W^2 Q)^{1/3} \cdot [c_f^{1/3} / j_H] \quad . \tag{2.42}$$

It is instructive to rewrite this in terms of normalized thermal resistance $R_{turb} = LW\theta$ and normalized hydraulic power $Q'' = Q/LW$; this gives

$$R_{turb} = LW\theta > (\mu^2 / k_c^2 \rho^2 C Q'')^{1/3} \cdot [c_f^{1/3} / j_H] \quad . \tag{2.43}$$

Note that this result is independent of the channel length. Now compare this with Eq. (2.38), which was obtained for laminar-flow high-aspect-ratio heat sinks (we have substituted $\Phi = 24$, $Nu = 8.235$):

$$R_{lam} = 4.37(\mu / k_c k_w^2 \rho^2 C^2 Q'')^{1/5} \cdot L^{2/5} \quad .$$

It is evident that since $R_{turb} \propto (Q'')^{-1/3}$ and $R_{lam} \propto (Q'')^{-1/5} L^{2/5}$, then for any given length L, there exists a critical mechanical pumping power flux

$$Q''_{crit} = (1.57 \times 10^{-5})[k_w^6 C \mu^7 c_f^5 / k_c^7 \rho^4 L^6 j_H^{15}]^{1/2} \quad ,$$

below which a high-aspect-ratio laminar-flow design will be superior to **any** flat-plate, high-Nusselt-number turbulent-flow design. For $Q'' > Q''_{crit}$, an optimized turbulent-flow design

may possibly be superior, but the actual crossover point where this occurs will probably be greater than Q''_{crit} because Eq. (2.38) is an inequality, and because we have completely neglected header pressure drops which turn out to be quite significant in turbulent-flow designs. In any case, the ratio R_{lam}/R_{turb} increases only as $(Q'')^{2/15}$, which is a very weak dependence.

To evaluate Q''_{crit}, we turn to experimental data for the values of c_f and j_H. For smooth annular gaps (parallel plates are an annular gap of infinite radius), over the Reynolds number range from 6,000 to 77,000, the Colburn factor may be approximated as $j_H \simeq 0.023 Re^{-0.2}$ [42], and the friction factor as $c_f \simeq 0.079 Re^{-0.25}$. Thus we have $Q''_{crit} = (0.2\ W\text{-cm}) L^{-3} Re^{0.875}$. But for the parallel-plate design to be competitive we also require that $Nu > k_w/k_c = 250$, which from Fig. 2-8 implies that $Re > 6 \times 10^4$, whence $Q''_{crit} \simeq 3000\ W/cm^2$ for $L = 1$ cm. That is, a pumping power (per unit of heated area) greater than 3000 W/cm^2 would be required to compete with our laminar-flow design, for a 1-cm channel length!

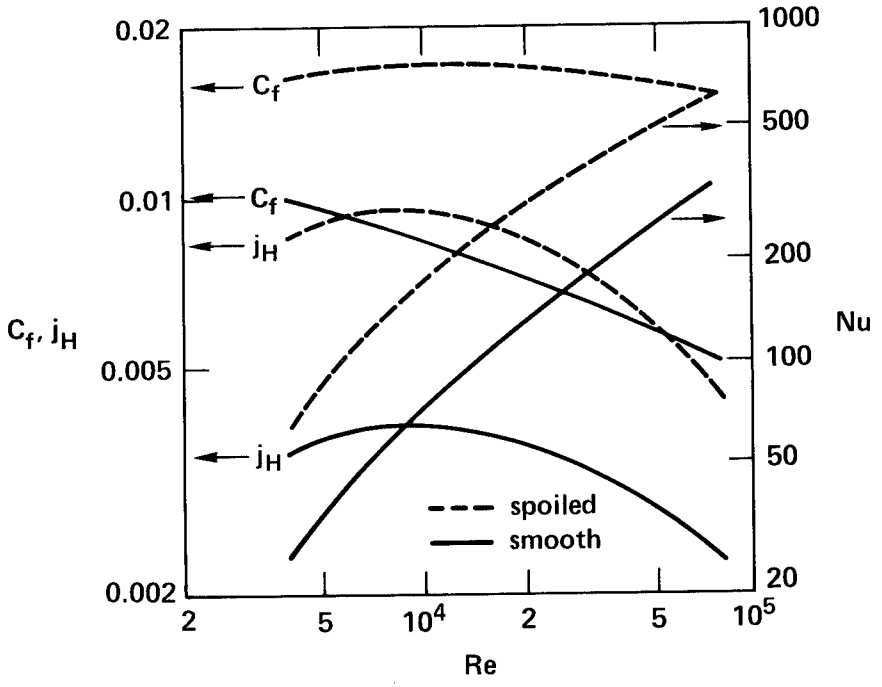


Figure 2-8: Nusselt number Nu, Colburn factor j_H, and friction factor c_f vs. Re for smooth tubes and for optimally roughened (spoiled) surfaces (from Ref. [40]).

Q''_{crit} can be reduced somewhat by the use of turbulence promoters ("vortex generators" or "surface spoilers") on the heated plate. Fig. 2-8 shows how carefully

roughened tubes can substantially increase Nu for a given Re. The friction-factor and Colburn-factor curves are quite flat, with the lowest (best) value of $c_f^{1/3}/j_H$ occurring around $Re \simeq 10^4$, where $c_f = .017$ and $j_H = .009$; thus $\dot{Q}''_{crit} \simeq 5$ W/cm^2 for L = 1 cm. This is a much more reasonable pump-power figure and is in fact approximately where we have been designing. Thus a well-designed, carefully roughened parallel-plate structure can possibly give performance equivalent to optimized laminar-flow silicon heat sinks, if the pump power exceeds 5 W/cm^2. If copper is used as a substrate material (rather than silicon), the break-even point is $\dot{Q}''_{crit} > 100$ W/cm^2. Thus a finned laminar-flow heat sink made of copper would clearly be preferable to a turbulent-flow design in practical applications if performance is the sole criterion; the laminar design would also be somewhat more compact.

Fig. 2-9 plots the lower bounds on normalized thermal resistance for the smooth-wall and rough-wall turbulent cases ($Nu \gg k_w/k_c$), as well as the exact result for the optimized laminar-flow case, assuming 25°C water. Note that whereas the rough-wall turbulent flow is essentially the best one can do with turbulent flow, one can keep reducing R in the laminar-flow design by reducing channel length L (hence reducing the spacing between headers). It should be noted that there is an intermediate design region in which $Nu \simeq k_w/k_c$; here there is **some** advantage to finning the surfaces, but the prime surface is also very important to heat transfer. In this regime, one can do slightly better than either the laminar or the turbulent curves in Fig. 2-9. For example, a very carefully optimized klystron heat sink design [20, 43] yielded R = 0.1 cm^2-°C/W for a local pump power of $\dot{Q}'' = 1.8$ W/cm^2.

To summarize our conclusions about flat-plate turbulent flow vs. finned laminar flow designs:

1. $Nu_T \gtrsim k_w/k_c$ is required to get significant potential advantages from turbulent flow.

2. At these high Nusselt numbers, we have calculated a lower bound on R which is not far from the best achievable performance by assuming c_f and j_H to be constants.

3. As pump power is increased, the performance of turbulent-flow designs will eventually overtake the laminar-flow design (for fixed L), but for $L \leq 1$ cm this occurs at very high pump powers $\dot{Q}'' > \dot{Q}''_{crit}$, which are not likely to be used in practical computer systems.

4. For any value of pumping power $\dot{Q}''$ per unit area, one can always design a laminar-flow heat sink with better performance than all possible turbulent-flow designs by choosing the

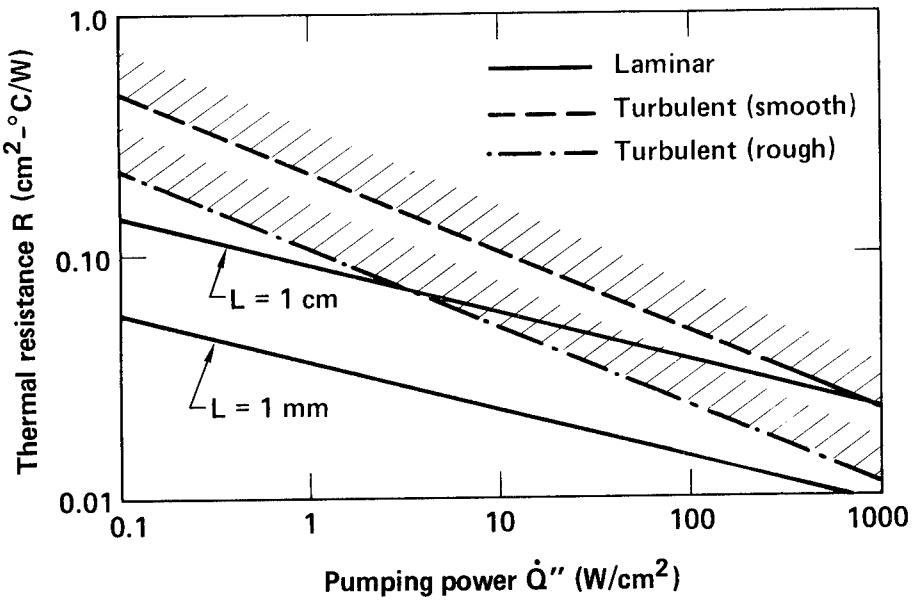


Figure 2-9: Optimized thermal resistance R (normalized for 1 cm^2) as a function of normalized pumping power $\dot{Q}''$ for laminar flow (L = 1 cm or 1 mm), and **lower bounds** for highly turbulent flow with smooth or optimally roughened pipes.

channel length L to be sufficiently small; if the substrate length is greater than L, several such heat sinks located end-to-end would be needed to cover the full substrate (Fig. 2-10).

5. Finally, it should be noted that we have completely neglected the header pressure drops associated with feeding the turbulent-flow heat sink. For a given channel length L, these pressure drops are usually relatively more severe than in the laminar-flow heat sinks, for which the manifold pressure contributes only a small fraction to the total pressure drop. The reason is that the turbulent-flow designs have wider channels, hence the length-to-width ratio is less, implying the header and entrance regions are relatively more important. Specifically, $P_{headers} = K\rho v^2$ (where K is <u>at least</u> of order unity), whereas $P_{core} = (L/H)c_f\rho v^2$; thus $P_{headers}/P_{core} \simeq H/Lc_f$. Typically $c_f \simeq 0.01$ for turbulent flow; thus $P_{header}/P_{core} \simeq 100(H/L)$, which can be of order unity or larger, and which increases as L is reduced. That is, the pressure drops associated with headering the turbulent-flow designs will greatly degrade their performance, and the degradation is increasingly severe as channel length L is scaled down. In contrast, the laminar-flow designs improve ($R \propto L^{2/5}$) as L is scaled down, and the header losses never play a significant role, as shown in the next section.

2.2.3. Friction-Coefficient Corrections

In the elementary optimization procedure, we calculated the pressure drop in a heat sink to be $P = (2L/D)(\rho v^2)c_{fm}$, where the laminar momentum boundary layer was assumed to be fully developed. That is, $c_{fm} = \Phi/Re$, where Φ is a **constant** determined from Fig. 2-5. For finite $L/(D \cdot Re)$, as in our designs, there will be an additional pressure drop ΔP_{entry} associated with the development of the momentum boundary layer. Shah and London [34] express this entrance effect in terms of a friction loss-factor K:

$$\Delta P_{entry} \equiv (\rho v^2/2)K_{entry} .$$

K_{entry} is a function of channel length, but for $L/(D \cdot Re) > 0.05$, it is essentially a constant (denoted K_∞). K_∞ is plotted in Fig. 2-5 and is approximately equal to 0.9 for the aspect ratios used in our designs.

Additional pressure drops will be associated with the input and output manifolds, or "headers". Kays [44] has calculated the irreversible entrance and exit loss factors for the case of a flat-duct heat exchanger having an abrupt-contraction entrance and an abrupt-expansion exit. In Kays' analysis, the pressure drop is $\Delta P_{header} = (\rho v^2/2)(K_e + K_c)$, where $K_e + K_c = 0.76$ for the 2:1 change in flow area used in our designs.

The entry-length effect and the header losses both contribute pressure drops proportional to the velocity head $(\rho v^2/2)$, hence they are experimentally indistinguishable. From now on we shall lump them together into a single loss factor K. Although calculations predict that $K \simeq 1.6$, in fact we have measured K to be between 3 and 4.5, probably because the header geometry is not as ideal as in Kays' calculations (see Section 3.2.3). Recalling that the core (channel) pressure drop was $P_{core} = (4L/D)(\rho v^2/2)(\Phi/Re)$, we have

$$P_{total} = P_{core} + (\rho v^2/2)K = P_{core}(1 + [K/4\Phi] \cdot [D \cdot Re/L]) . \qquad (2.44)$$

In our elementary designs, we found $D \cdot Re/L = 6(Nu/Pr) \simeq 8$ (for constant power), or $D \cdot Re/L = 12(Nu/Pr) \simeq 20$ (for constant pressure). Estimating $K \simeq 3$ leads to a net pressure increase of 28% in the constant-pumping-power design, and 62% in the constant-pressure design. These increases are large enough that it is worth reoptimizing the designs to take into account the effects.

2.2.3.1. Constant-Pressure Case

It is convenient to analyze the problem by using $\chi \equiv D \cdot Re/L$ as an independent variable, rather than D. The two are related by noting that $\chi = vD^2\rho/\mu L$ and $P_{core} = 2\mu\Phi Lv/D^2$, whence

$$D = (2\mu^2 L^2 \Phi/\rho P_{core})^{1/4} \chi^{1/4} .$$

Using Eq. (2.44) to substitute in for P_{core}, we have

$$D = (2\mu^2 L^2 \Phi/\rho P)^{1/4} [\chi(1 + K\chi/4\Phi)]^{1/4} . \qquad (2.45)$$

The analysis of Section 2.1.2.1 may then be carried through as before. Substituting for D into Eqs. (2.28) and (2.29) leads to the conclusion that θ is minimized when

$$\chi(1 + K\chi/4\Phi) = 12(Nu/Pr) \simeq 20 .$$

This quadratic equation in χ is readily solved when K is known; for example, $K = 3 \Rightarrow \chi \simeq 14$. The thermal resistance is increased by a factor $(1 + K\chi/4\Phi)^{1/4} \simeq 1.093$ in this example; the optimal channel dimension D is increased by the same factor.

2.2.3.2. Constant-Pumping-Power case

This case may be analyzed in much the same way. The result is

$$\chi(1 + K\chi/4\Phi)^{1/2} = 6Nu/Pr \simeq 8 .$$

For $K = 3$, we have $\chi \simeq 7.2$. The thermal resistance and optimum channel dimension are both increased by a factor $(1 + K\chi/4\Phi)^{1/5} = 1.046$. Note that the constant-pumping-power optimization is much less affected by these entrance/exit losses than is the constant-pressure case, because its design results in a smaller value of χ.

2.2.4. Nonlinearities in Thermal Resistance

In our analyses, the thermophysical parameters of the coolant (ρ, C, k_c, and μ) were considered to be constants; in real materials the parameters vary with temperature, causing nonlinearities in the thermal resistance. For most coolants of interest, the variations in ρ, C, and k_c are small and can be ignored. For example, the density of water decreases by only 4% from 0°C to 100°C ($\rho_{0°C} = 1.000$ gm/cm^3; $\rho_{100°C} = 0.958$). Its heat capacity varies by less than 1% over that range. Water's thermal conductivity increases by 21% ($k_{0°C} = 5.61 \times 10^{-3}$ W/cm-K; $K_{100°C} = 6.81 \times 10^{-3}$), which means our design is slightly conservative (we used the 20°C value). To make the design more precise one could use the thermal conductivity at the expected output water temperature $T_{output} = T_{input} + \theta_{cal}\dot{Q}$, since we desire to minimize the temperature at the hottest (downstream) location on the substrate. This typically leads to a 1 or 2% improvement in the optimum heat transfer when operating at a 100°C peak substrate temperature.

In contrast, the viscosity of most liquids varies dramatically with temperature; for example, the viscosity of liquid water decreases more than six-fold from 0°C to 100°C. The flow velocity profiles at any particular cross section of a heated channel will be modified because of the temperature variation, resulting in a lower local friction factor that would be expected based on the mean water temperature. For circular tubes, Deissler [45] has found an approximate correction factor $(\mu_w/\mu_m)^{0.58}$ which accounts for this effect. Here μ_w is the viscosity at T_w (the wall temperature) and μ_m is the viscosity at T_m (the mean coolant temperature at the cross section in question). Assuming this result holds for our high-aspect-ratio channels, our optimization formulas should therefore be corrected by replacing Φ by $(\mu_w/\mu_m)^{0.58}\Phi$, where μ_w and μ_m are the viscosities evaluated at temperatures halfway downstream. The local friction factor will itself vary along the length of the channel due to the viscosity variation, hence the total pressure drop will be affected. This effect can be incorporated by evaluating the water viscosity at the mean coolant temperature halfway downstream, i.e., μ evaluated at $T = T_{input} + \theta_{cal}\dot{Q}/2$. The heat transfer will also be improved by the change in cross-sectional velocity profile induced by the viscosity variation. For laminar flow in a circular tube, Deissler [45] suggests that Nu be corrected by the factor $(\mu_w/\mu_m)^{-0.14}$.

For water cooling, we can estimate the maximum benefit in thermal resistance due to these effects by assuming the peak wall temperature (i.e., the surface temperature) is 100°C at the downstream end and the input water temperature is 20°C. For the constant-pumping-power optimization, we would have the flow friction (Φ) decreased by a factor $(\mu_{84°C}/\mu_{36°C})^{0.58}(\mu_{36°C}/\mu_{20°C}) = 0.4601$, and the effective Nusselt number increased by a factor $(\mu_{84°C}/\mu_{36°C})^{-0.14} = 1.1086$. The optimum thermal resistance would thus change by the factor $(0.4601/1.1086)^{-1/5} = 0.839$, a significant 19% increase in heat transfer. A similar calculation for the constant-pressure case yields a factor of $[(\mu_{90°C}/\mu_{30°C})^{.58+.14}(\mu_{30°C}/\mu_{20°C})]^{1/4} = 0.799$ reduction in θ. The channel geometry must be appropriately modified to achieve this optimum, of course.

Note that the viscosity reduction with increasing temperature actually makes the fins more effective than our calculations have indicated. The bases of the fins, being hotter, will experience a higher flow velocity than the upper portions. Hence the temperature gradient in the z-direction will not be quite as large in the coolant as would be the case for a constant-velocity flow field.

If any portion of the channel surface exceeds the saturation temperature of the coolant, vapor bubbles will form on the surface. However, since the bulk of the fluid is presumably

below the boiling point, the phenomenon of "subcooled boiling" will occur in which the bubbles grow only until they reach the subcooled region and then collapse, as discussed by London [20]. This process of bubble formation and collapse can enhance the convective heat transfer coefficient h. From experimental evidence, London speculates that h might increase by 50%; he suggests that thermal runaway (burnout) will occur when 7-10% of the surface is covered by bubbles. He notes that it is very desirable to eliminate sharp corners, as the stagnant boundary layer would result in more bubble coverage (and hence early burnout) in corners. Our fabrication techniques fulfill this criterion; microchannels produced using KOH etchant have no acute angles (Fig. 3-7). The actual benefit of the increased heat transfer would be rather minimal in our designs, because the effective area enhancement factor due to fins is $\alpha_c = 5.43$, whereas the enhancement in heat transfer due to subcooled boiling only occurs at the "primary surface" at the bottom of the grooves. Thus a 50% enhancement in h at the groove bottom would only represent approximately a 10% increase in total heat transfer.

Estimates have been made of the heat flux which one may achieve in the subcooled boiling regime before burnout occurs, but they are empirically derived and may not apply to fully-developed laminar flow. The main point is that the subcooled boiling phenomenon allows one to exceed a surface temperature of 100°C and simultaneously to achieve a modest improvement in heat transfer coefficient, provided one does not exceed the burnout flux.

The final nonlinearity in θ is due to the variation in the substrate thermal conductivity k_w. Whereas the thermal conductivity of metals is quite insensitive to temperature at normal temperatures ($k_{Cu} = 4.01$ W/cm-K@0°C; $k_{Cu} = 3.93$@100°C), the thermal conductivities of insulators or semiconductors are quite temperature-sensitive. As shown in Fig. 1-2, silicon's thermal conductivity is $k_{Si} = 1.48$ W/cm-K at 27°C, 1.19 at 77°C, and 0.99 at 127°C. We can attribute 40% or 50% (for constant P or constant $\dot{Q}$, respectively) of the total thermal resistance of an optimized heat sink to fin conduction. At the downstream end, the average fin temperature may be thought of as being 20 or 25% below the maximum temperature rise, i.e., 84°C or 80°C if $T_{input} = 20$°C and $T_{max} = 100$°C. $k_{Si} \simeq 1.17$ W/cm^2-K at this temperature (vs. 1.52 in our simple design), which implies that the optimal thermal resistance will increase by a factor 1.140 (constant P) or 1.114 (constant $\dot{Q}$). This nullifies much of the viscosity advantage discussed earlier.

To summarize, several nonlinear effects result from large power fluxes, due to variations in material properties with temperature. The net effect is a slight improvement in thermal

resistance $\theta_{nonlin} = \Delta T/\dot{Q}$ over its linear value θ_{linear}; the changes are summarized in Table 2-3 for a 100°C maximum surface temperature and a 20°C input temperature. To achieve these changes, the design channel dimension D should be reduced by 14.6% (in the constant-$\dot{Q}$ case), or 18.9% (in the constant-P case).

Constant	θ_{linear}	Effect of $\mu(T)$	Effect of $k_c(T)$	Effect of $k_{Si}(T)$	Net effect	θ_{nonlin}
$\dot{Q}$	.063°C/W	−16.1%	−1.5%	+11.0%	−8.3%	.055°C/W
P	.055°C/W	−20.1%	−1.2%	+14.0%	−10.0%	.045°C/W

Table 2-3: Effects of nonlinear material parameters on optimized θ (water-cooled Si).

2.2.5. Thermal Spreading in the Silicon Substrate

Our analysis has thus far neglected any conductive thermal resistance $\theta_{cond} \equiv R_{cond}/LW$ between the planar heat source and the finned heat sink. In fact, there will be an additional thermal resistance associated with the bulk, unetched residual silicon between the bases of the fins and the heat source (the distance $H_s \equiv t_{Si} - H$ shown in Fig. 2-2). If $H_s \gg w_c$, then $R_{cond} \simeq H_s/k_w \propto H_s$. On the other hand, if $H_s \ll w_c$, then the heat generated beneath a channel must conduct through a relatively narrow gap of width H_s before reaching the fins; $R_{cond} \propto H_s^{-1}$ in this case, where the constant of proportionality will depend on the exact geometry of the channel bottoms; it could be calculated by conformal mapping. Note that the angled or rounded bottoms normally obtained by orientation-dependent etching or precision sawing (Figs. 3-3 and 3-4) are much more favorable from a thermal spreading viewpoint than are the flat bottoms sketched in Fig. 2-2. Clearly an optimum value of the residual silicon thickness H_s exists which minimizes R_{cond}. While the exact result depends on the shape of the channel bottoms, it is reasonable to approximate the minimum possible thermal resistance as $R_{cond} \simeq w_c/k_w$, where H_s would be slightly less than w_c to achieve this value. Referring to Eq. (2.31), we see that $R_{cond}/R_{opt} \simeq 3k_c Nu\alpha_c/8k_w = 0.069$; hence we expect the conductive thermal resistance of the residual silicon to add an additional 7% to the thermal resistance of an optimized micro-heat sink. This 7% is for the worst case, i.e., directly over the center of a channel. Over the fins, R_{cond} will be slightly less, so a slight spatial variation in temperature (having the same period as the fins) might be observed.

It should be noted that a residual silicon thickness $H_s \simeq w_c$ is perfectly adequate from a structural mechanics viewpoint. Since the aspect ratio of this region is approximately unity, a

coolant pressure comparable to the fracture stress of bulk silicon (σ_{Si} = 347 MPa = 3420 atm) [46] would be required to fracture it.

We have also neglected lateral thermal spreading in our analyses, i.e., heat conduction in the x-y (substrate) plane. This is only important near the perimeter of the heat source, where the generated heat flux abruptly transitions to zero. Whereas our model would predict an abrupt drop in temperature at this transition, lateral heat conduction will smooth out the transition. This will be important experimentally, for it means that the maximum surface temperature will occur somewhat short of the downstream end of the heat source. Moreover, the maximum temperature will be slightly less than predicted by our simple model.

We can estimate the magnitude of this effect by modelling the conducted heat flux in the x-direction as a uniform flux J_x occupying a depth H_{eff} into the silicon substrate. Thus

$$\dot{q}(x) = h_{IC}\triangle T(x) + H_{eff}\partial J_x/\partial x\,, \tag{2.46}$$

where $\triangle T$ is the substrate temperature rise above the coolant temperature, h_{IC} is the local heat-transfer coefficient at the substrate surface, and $\dot{q} = \dot{q}_o u_{-1}(L - x)$ is a step function. J_x is assumed to be driven by Fourier's law:

$$J_x = -k_{avg}\partial(\triangle T)/\partial x \tag{2.47}$$

where k_{avg} is an average thermal conductivity of the fin structure. k_{avg} accounts for the fact that the heat sink is not solid silicon; we'll approximate $k_{avg} \simeq k_w/2$ for our designs ($w_c = w_w$). Eqs. (2.46) and (2.47) can be solved to yield:

$$\triangle T(x) = \begin{cases} (\dot{q}/2h_{IC})[2 - e^{(x-L)/L_o}] & \text{for } x<L \\ (\dot{q}/2h_{IC})[e^{(L-x)/L_o}] & \text{for } x>L\,. \end{cases}$$

Here $L_o \equiv \sqrt{k_{avg}H_{eff}/h_{IC}}$ is the decay length of this lateral thermal spreading. Taking $h_{IC} = k_c Nu\alpha_c/2w_c$, $H_{eff} \simeq 2\alpha_c w_c$, and $k_{avg} \simeq k_w/2$, we have $L_o \simeq (2k_w/k_c Nu)^{1/2} w_c = (0.71)H_{eff}$. Thus the lateral spreading distance is the same order as the effective depth of the heat sink, an intuitively reasonable result. For a heat sink having $H_{eff} = 500\ \mu m$ and a heated area of length $L = 1$ cm, such as was used in our experiments, we predict that the maximum temperature T_{max} will occur 1.3 mm short of the downstream end of the heater, at which point T_{max} is 4% less than the predicted maximum.

A similar thermal spreading situation would exist in the y-direction, except that now

$$H_{eff}k_{avg} \simeq H_s k_w + (2\alpha_c w_c)k_c Nu/2)$$

for heat flow in the y-direction (across the fins, rather than along them). Thus $L_o \simeq \sqrt{H_s H_{eff} + 2w_c^2}$. In our typical experiments, $H_s = 200\ \mu m$, $H_{eff} = 300\ \mu m$, and $w_c = 50\ \mu m$; hence $L_o \simeq 250\ \mu m$, i.e., somewhat less lateral thermal spreading occurs in this direction.

2.2.6. Ultimate Limits

In this section some limits to further heat-transfer improvements are discussed. We have previously pointed out that for a fixed pumping power/unit area ($\dot{q}$"), the heat transfer from a surface can be increased by scaling down the channel length L; $R \propto L^{2/5}$ as shown in Eq. (2.38). None of our design refinements interfere with this scaling, because all the corrections were percentage changes in thermal resistance; hence they scale down with R. Specifically, the header pressure drops and the thermal spreading resistance both remain small perturbations in the scaled designs. For substrates which are longer than the desired channel length L, one would have to integrate a series of heat sinks end-to-end as sketched in Fig. 2-10 (multiple coolant feed points are required).

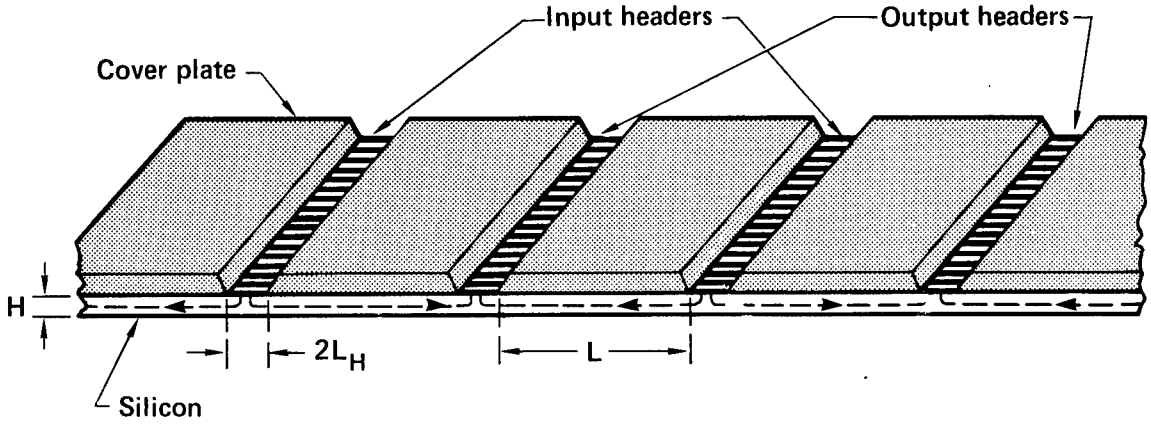


Figure 2-10: A multiple-header arrangement to allow scaling down of the channel length L. The header width L_H should be comparable to the silicon thickness H for proper heat transfer over the header regions.

The question remains whether suitable heat transfer could be achieved in the areas over the headers. The answer appears to be yes, provided the headers are properly designed. Specifically, the length L_H of the header (in the direction of flow) should be roughly equal to the channel depth H. If L_H were much smaller than H, then the fluid would be squeezed through a smaller flow area than the heat sink itself, resulting in a much larger pressure drop than was calculated in Section 2.2.3. The heat transfer would be excellent, however. If L_H

were much larger than H, there would be a stagnant region having poor heat transfer because most of the fluid would flow near the inner portion of the header, over a region of length $\simeq$H in order to minimize channel pressure drop. Even if the header areas had relatively poor heat transfer, the thermal spreading length L_o (discussed in Section 2.2.5) is comparable to H, hence would prevent those areas from getting excessively hot provided $L_H \leq L_o \simeq H$. Finally the cover plate thickness should scale down with H, to prevent the flow friction in the headers from becoming excessive. The only limit to scaling down L occurs when the channel length L becomes comparable to the channel depth H; at this point the entire flow problem becomes three-dimensional and it is not at all clear what happens. Moreover header resistance becomes comparable to core flow resistance, so we can no longer speak of the header and core as being separate. For typical operating pressures, this limit is reached when L$\simeq$100 μm, by which point a 6-fold reduction in thermal resistance is predicted.

Another way to improve heat transfer is the brute-force approach of greatly increasing the pumping power flux $\dot{q}''$ or pressure P; the heat transfer scales as $(\dot{q}'')^{1/5}$ or $P^{1/4}$. The only fundamental limit here occurs when the viscous heating $\Delta T_{visc} = P/\rho C$ of the coolant becomes significant. For water, $\Delta T_{visc} = (0.024^{\circ}C/atm) \cdot P$, hence a 1000-atmosphere pump pressure would contribute a temperature rise of 24°C due to viscous heating, about as high as would likely be tolerable. This is also a maximum reasonable pressure from a structural viewpoint too. The associated performance improvement would be about 5-fold over our present designs.

Combining the above procedures (increasing P to 1000 atm and scaling L down to 100 μm), we could conceivably remove 50 kW/cm^2 from a silicon substrate while maintaining substrate temperatures below 120°C. However it seems unlikely that VLSI circuits would be designed to exceed power densities of 1000 W/cm^2 over large areas because the thermal spreading resistance from individual devices could become limiting. Thus the extraordinary efforts just discussed are probably not worth the effort to implement.

Chapter 3
Microscopic Silicon Heat Sinks: Experiments

3.1. Fabrication

3.1.1. Silicon Micromachining

In order to fabricate optimized heat sinks in silicon substrates, fabrication techniques capable of making microscopic high-aspect-ratio channels with great precision are required. The need for precision in manufacturing has been pointed out by Shah and London [47], who showed that manufacturing variations of the order of 10% in laminar-flow heat exchangers could cause a significant (the order of 25%) reduction in heat transfer with negligible benefit to the flow friction. Since the dimensions involved are considerably smaller than those used in conventional compact heat exchangers [17], the use of integrated-circuit microfabrication techniques was explored. Two such techniques have been successfully used in this work: orientation-dependent etching and precision mechanical sawing.

3.1.1.1. Orientation-Dependent Etching

It is known that certain etchants for crystalline silicon are very sensitive to crystallographic orientation. Specifically, etchants such as ethylene diamine pyrocatechol (EDP) or potassium hydroxide (KOH) etch the <111> planes of silicon very slowly in comparison to other directions [48]. In effect, the etching process stops at any <111> plane, or at any concave ($\theta<180^{o}$) intersection of <111> planes. A silicon wafer whose surface is a <110> plane will have two sets of parallel <111> planes oriented at right angles to the surface; the angle between the two sets of planes is 70.53^{o}. (Two additional sets of planes exist which subtend a shallow 35.26^{o} angle with the wafer surface; these are not of interest here.) If the silicon surface is masked by a series of parallel stripes which are perfectly aligned with one of these sets of <111> planes, then in an aqueous solution of 50% KOH the unprotected stripes will etch vertically downward into the silicon with almost no lateral etching (i.e., no undercutting). Kendall [49] has shown that aspect ratios of greater than 600:1 can be achieved when the mask is perfectly aligned with the <111> planes, which is far in excess of that needed for the fabrication of optimized micro-heat sinks.

In order to get good, defect-free results with orientation-dependent etching using KOH, a suitable masking material must be used. Photoresist will not withstand KOH, but SiO_2 or Si_3N_4 will. Plasma-deposited nitride was found to have pinholes, so we used either thermally-grown or sputtered SiO_2 for masking against KOH; the SiO_2 itself is patterned using standard photolithographic techniques. Unfortunately SiO_2 is slowly attacked by KOH; Fig. 3-1 shows Kendall's data [50] for the etch rates for SiO_2 and for unobstructed <110> silicon as a function of temperature (the etch rate of narrow <110> grooves is about 70% of the unobstructed rate). Note that the selectivity improves as temperature is reduced. For etching 300-μm deep microchannels, a temperature of 52°C and a SiO_2 mask thickness of 1.1 μm was found to be adequate; etching time was typically 25 hours. Etching at higher temperatures often resulted in defects, even with a nominally adequate oxide mask thickness.

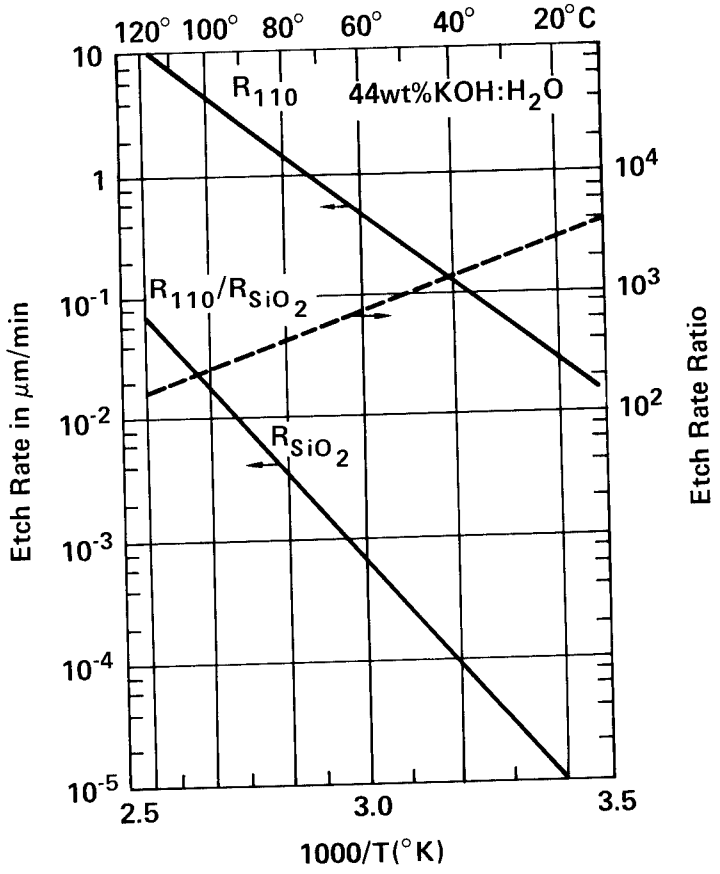


Figure 3-1: Etch rate of unobstructed <110> silicon in KOH. Narrow grooves etch at 70% of these values. Also shown are maximum SiO_2 etch rates (data of Kendall [50]).

It is often necessary to mask one whole side of a wafer while etching the other side in KOH. If it is not possible to use SiO_2, then certain waxes may be used provided the KOH

temperature does not exceed 50°C. Alternatively, Dynatex "Wafer-Grip" [51] (a hydrocarbon film) will, if properly applied, withstand hot KOH for many hours before starting to peel.

The effect of misalignment of the mask pattern with respect to the <111> planes is an increase in the amount of undercutting which occurs; this has been explained by Kendall as due to etching of the misorientation ledges [49]. He found that the amount of undercutting from each edge is approximately $\theta H/35$, where θ is the misorientation angle in degrees and H is the channel depth. Interestingly, the structures are still nearly vertical-walled (not tapered) despite the undercutting. The only irregularities are due to "misorientation ledges", which cause an occasional jump of 1/2 μm or so in the channel width. Such microscopic irregularities are believed to be too small to affect the heat-transfer or flow-friction characteristics of the channels at the low Reynolds numbers which we will be operating. Because the walls are essentially vertical even with a substantial undercut of the oxide, one may vary the fin/channel (w_w/w_c) ratio simply by misaligning the wafer, rather than fabricating a large set of masks. Of course, the period $w_c + w_w$ cannot be changed except by using a different mask.

The wafer "flat" which is used for alignment is generally only accurate to within $\pm 1^\circ$, hence for our microchannel design depths of $\sim$400 μm, a total widening of the vertical groove width of up to 23 μm may occur if the flat is used as the sole alignment aid. Since this is a large and unpredictable error relative to the design channel width of $\sim$50 μm, an initial etch test was normally performed near the edge of each wafer using a "splay" pattern. This pattern has a series of 5-μm wide lines, each rotated by 0.1° relative to its neighbor. By etching the splay pattern to a depth of 100 μm, the width of adjacent lines differs by about 0.6 μm, which is easily measurable in an optical microscope having a reticle eyepiece. It is often difficult to identify visually the absolute minimum width groove due to mask dissolution and linewidth variations, so the procedure we used was to graph the etched line width as a function of angle on both sides of the true <111> plane. A straight line was fit to the data on each side, and the intersection of the lines defines the orientation of the <111> plane. An edge of the wafer was then diced off parallel to this angle to provide a long, very accurate flat for subsequent alignment on either side of the wafer.

The splay pattern itself only covers $\pm 5^\circ$ in our mask; thus one normally uses the wafer "flat" to align the splay pattern for the aforementioned etch test. Usually the flat is located in a $\langle 1\bar{1}0 \rangle$ direction, in which case one has <111> planes 35.26° in either direction from the flat. In other cases the flat is already aligned parallel with one of the <111> planes. In yet

another set of <110> wafers, the major flat was a $\langle\bar{3}3\bar{1}\rangle$ plane and the minor flat was a $\langle\bar{1}1\bar{2}\rangle$ plane. To unambiguously locate the <111> planes in a wafer where the flat location is unknown or suspect, we examined the etch pit produced by etching of a small hole in a thermally oxidized <110> wafer. The limiting shape of such an etch pit is always a hexagon as shown in Fig. 3-2; two pairs of edges show the location of the vertical $\langle 1\bar{1}1\rangle$ and $\langle\bar{1}11\rangle$ planes (70.52° apart), and the third pair identifies the two low-angle <111> and $\langle 11\bar{1}\rangle$ planes, which are not of interest here.

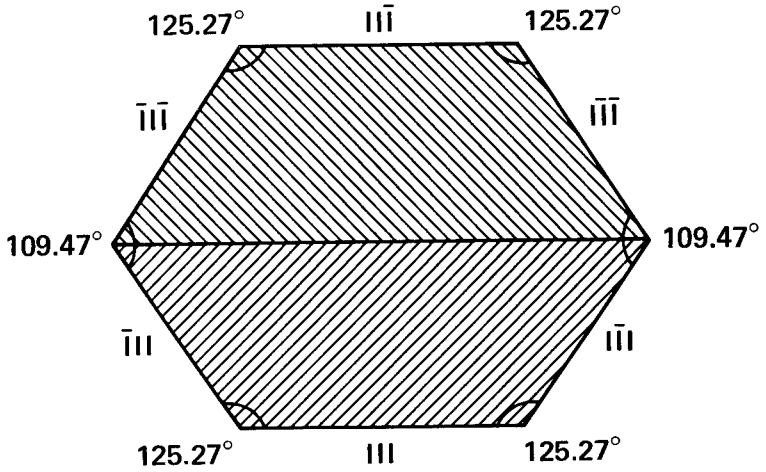


Figure 3-2: General shape of the etch pit formed when etching through a small hole in the SiO_2 mask covering a <110> silicon wafer using KOH. The pit is bounded on all sides by <111> planes, two pairs of which are perpendicular to the surface.

The procedure for etching was to prepare a fresh solution of 50% KOH and 50% H_2O (by weight) in a glass beaker. The solution temperature was allowed to stabilize in an oven to within 1°C of the desired temperature. The patterned wafers were then dipped into 50:1 HF, rinsed and dried, then immediately inserted into the KOH solution. If the HF dip is not performed, a thin SiO_2 layer will inhibit the etching for a minute or so until it is removed by the KOH; since the thin layer is not perfectly uniform in thickness, an irregular microchannel depth profile results. Fig. 3-3 shows electron micrographs of a typical microchannel array after etching in KOH; the oxide mask is visible as it overhangs the ends of the silicon "fins".

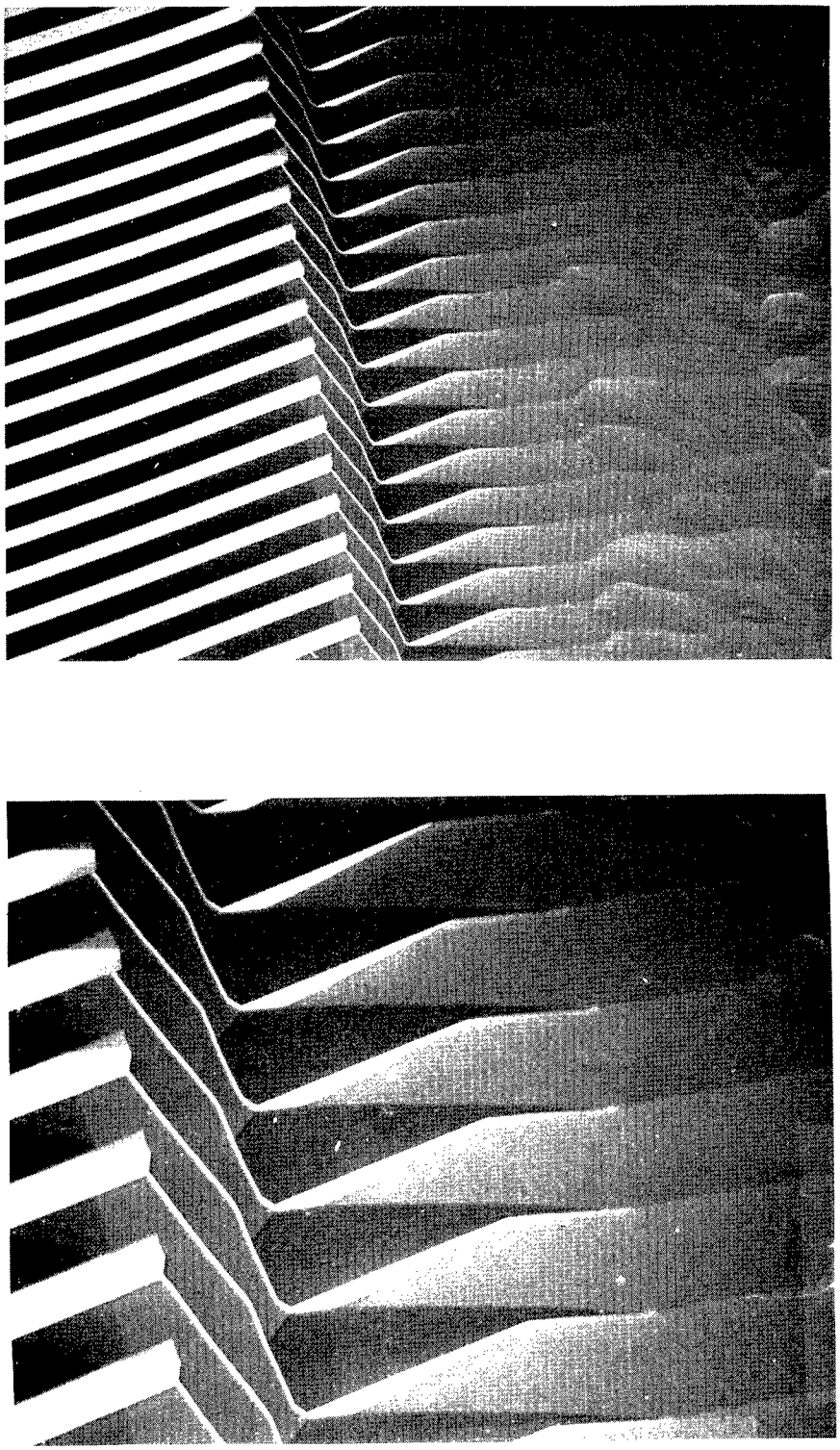


Figure 3-3: SEMs of microchannels etched in <110> silicon using KOH. The spatial period is 100 μm.

3.1.1.2. Precision Mechanical Sawing

Precision mechanical sawing is an alternative to orientation-dependent etching which was found to be suitable for creating high-aspect-ratio microchannels in hard materials. There are two situations in which precision sawing is preferable to ODE. First, when one desires to fabricate the microchannels in a substrate other than <110> silicon (for example, gallium arsenide). Second, precision sawing allows the fabrication of microscopic "pin-fin" structures, of which Fig. 3-4 is an example.

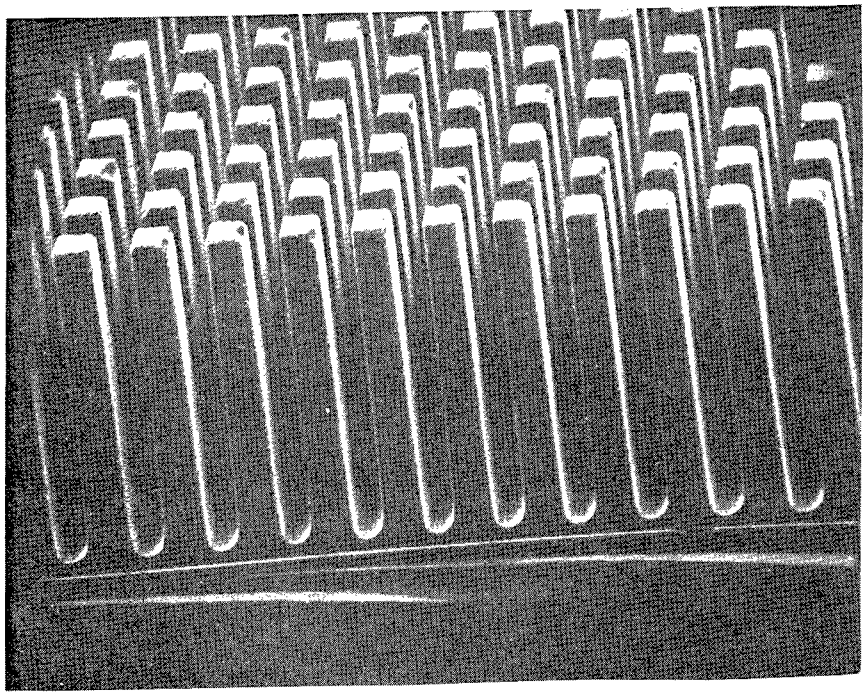


Figure 3-4: SEM of rectangular pin-fin structures fabricated in silicon by precision mechanical sawing. The spatial period is 80 μm in both directions.

Such structures cannot be fabricated by ODE, because the etching of **convex** surfaces such as pin-fins is determined by the fastest-etching planes, which means that the pin-fins will undercut very rapidly and aspect ratios greater than approximately unity are not obtainable. In contrast, the etching of deep, long microchannels expose only **concave** surfaces, and thus the slowest-etching planes (the vertical <111> planes) limit the etch rate, resulting in negligible undercutting. This distinction is discussed in detail in papers by Batterman [52], Jaccodine [53], and others. Precision mechanical sawing of high-aspect-ratio grooves is accomplished using a standard semiconductor dicing saw, in this work a "Tempress® 602". The cutting tool is a precision metallic blade with diamonds impregnated in its sides. The blade typically rotates at 25,000 RPM or faster, and is continually bathed in a jet of coolant.

The width and smoothness of the cut is determined by the metallic blade width and by the diamond grit size. A large variety of blade widths are commercially available; the narrowest produce 25-μm cuts. We generally used blades in the 40- to 60-μm range. The maximum possible depth of the cut, called the "blade exposure", is determined by mechanical considerations; typically aspect ratios of 20:1 are obtainable, which is well in excess of our requirements. The cut widths typically vary by no more than 2%.

In our work, the saw was programmed to cut parallel grooves of a preset depth in silicon and GaAs (typically 50 μm wide on 100-μm centers, with depths of 300 to 400 μm). To construct pin-fin arrays, the chuck was rotated 90^{o} and another set of perpendicular cuts was made. Proper lubrication and cooling of the blade is essential for minimizing chipping and preventing fracture of the fins. Very successful results were achieved using a lubricant additive (Dynatex Kerf-Aid®) in the blade's cooling water. Fig. 3-4 is an electron micrograph of a precision-sawn silicon pin-fin array having 40 μm$\times$40 μm$\times$400 μm pins on 80 μm$\times$80 μm centers. Such arrays were also successfully fabricated in GaAs. Following the sawing operation, a brief wet chemical etch of the substrate was usually done to remove any microcracks which may have begun to develop during the mechanical sawing operation [54].

The primary disadvantage of precision sawing is that, due to the large blade radius (typically 2.78 cm), one is restricted to fabricating long grooves which extend the entire length of the substrate wafer. Thus some extra demands are made on the designer to achieve suitable sealed packages (see Section 3.1.4). For this reason, it is likely that for manufacturing of practical micro-heat sinks which cannot be made using ODE, other techniques for fabricating the high-aspect-ratio plate-fins or pin-fins would be preferable to precision sawing. Examples might be reactive-ion etching, chemically-assisted ion milling, or sandblasting. Nonetheless, precision mechanical sawing proved to be a fast, convenient and practical fabrication technique for research purposes.

3.1.2. Bonding Materials to Silicon

In order to confine a coolant fluid to within the microchannels, some sort of cover plate must be bonded to the tops of the silicon micro-fins. The bonding technique must provide a leak-tight seal at normal operating pressures (up to 50 psi = 345 kPa) and temperatures (up to 100^{o}C). Furthermore, the thermal expansion coefficient of the cover plate should be well matched to that of the silicon; silicon would be the ideal choice for a cover plate. A technique for bonding silicon to silicon by reflowing a thin adhesive layer of phosphosilicate glass has

been reported in the literature [55]. However, the technique required a quartz vacuum chuck to clamp the wafers together while heating them to 900°C, which is difficult to fabricate.

Another possible technique for bonding silicon to cover plates is to use an epoxy resin. In order to achieve a sufficiently thin layer (so as not to clog the microchannels), we spun on the epoxy in the same way that photoresist is spun onto wafers. According to theory [56], the thickness t_{liquid} of a nonvolatile Newtonian spun-on liquid is $t_{liquid} = (3\upsilon/4\omega^2 t)$, where υ is the liquid's kinematic viscosity, ω is the angular spin frequency, and t is the duration of the spin. Thus a fairly low-viscosity epoxy (υ = 500 centistokes), spun at 4800 RPM for 60 seconds, will give a 5 μm layer, which is satisfactory. Fig. 3-5 is a cross section of a sawn silicon microchannel array bonded to a Pyrex cover plate using this technique.

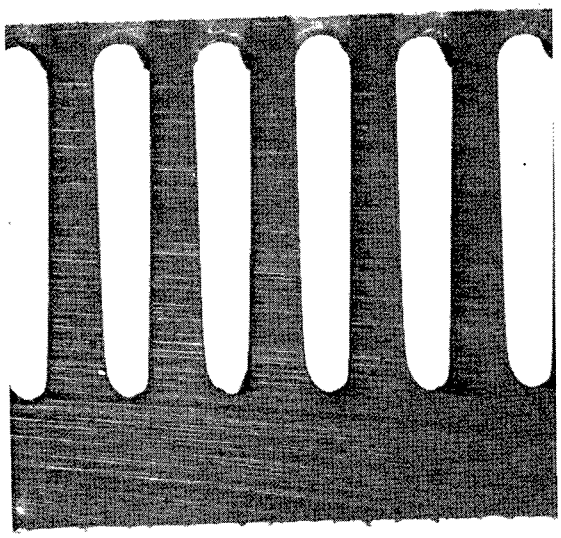


Figure 3-5: Transmission photomicrograph of a silicon microchannel heat sink cross section, where a spin-on epoxy adhesive was used to bond the cover plate. These channels were formed by precision sawing (100 μm period).

Note the hardened epoxy meniscus near the end of each fin, indicating that good wetting occurred before the epoxy cured. The tensile strength of this bond was satisfactory (>50 psi) at room temperatures, but in the presence of very hot water (90°C) the low-viscosity epoxies which we used would peel. Thus they were suitable for fluid-flow measurements but not for high-power heat sink designs; perhaps an epoxy suited for high-temperature operations would perform better.

The bonding technique which gave the most consistently satisfactory results was "anodic bonding", or "field assisted bonding", which is a well-established technique for bonding glass to silicon or to certain metals [57, 58, 59]. The glass and silicon are placed in contact and heated to about 400°C, at which temperature the glass is conductive and beginning to soften. A dc potential of several hundred volts is applied across the materials with the silicon as anode. The large electrostatic attraction, estimated to exceed 350 psi, possibly combined with resistive heating associated with the ionic current in the glass, results in a gradual fusion of the surfaces. Unlike the previously-described bonding techniques, anodic bonding does not employ an adhesive layer. It is therefore necessary that both surfaces be very clean, optically smooth, and very flat. A small entrapped dust particle will usually result in a substantial region of noncontact (0.5 mm or so) around it, far out of proportion to its size. Typically a relative flatness of 20 fringes (viewed under a sodium lamp) over a 2-inch area is sufficient for a complete seal between very clean surfaces.

Borosilicate glass such as Pyrex 7740 is commonly used when anodically bonding to silicon, because the thermal expansion mismatch is very small.

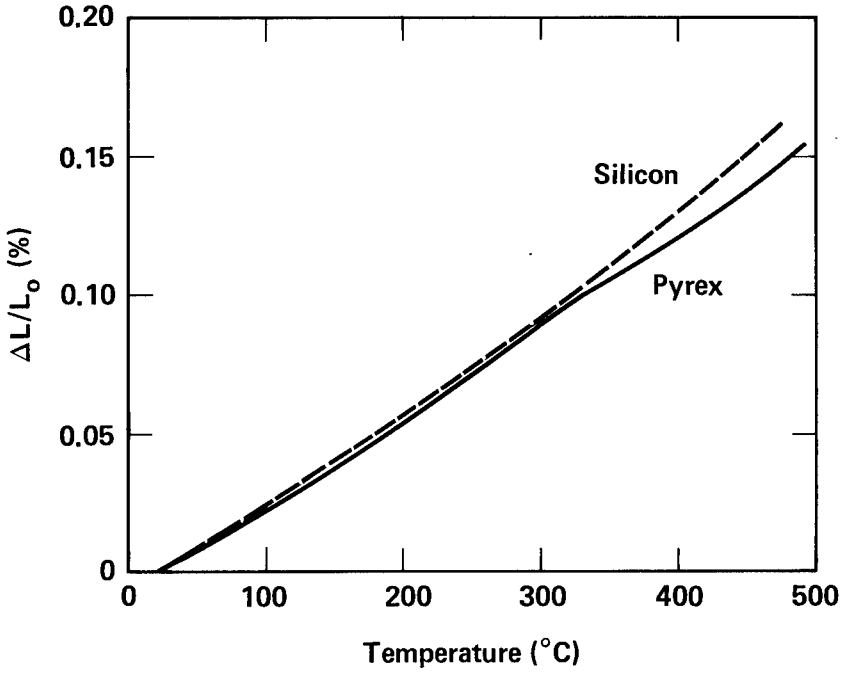


Figure 3-6: Linear thermal of expansion of silicon and Pyrex 7740 (from Ref. [60]).

Fig. 3-6 is a plot of the published linear thermal expansion of silicon and Pyrex, where 20°C is taken to be the zero point [60]. Up to 300°C, the match is nearly perfect (only ~20 ppm of expansion mismatch). As the bonding temperature is increased above 300°C the differential

expansion increases, causing the bonded assembly to bow slightly at room temperature (the silicon face is concave). This bowing was used to simulate wafer warpage when testing the microcapillary thermal interface concept in Chapter 5.

A proper anodic bond provides a strong, reliable, quite hermetic seal; it was extensively used in the fabrication of micro-heat sinks. Fig. 3-7 is an optical micrograph of a cross section of a silicon microchannel array (fabricated by ODE etching) which has been anodically bonded to a Pyrex cover plate. Fig. 3-8 is an electron micrograph of the array, viewed 45^{o} from the normal. The chipping is due to the use of a dicing saw to section the heat sink.

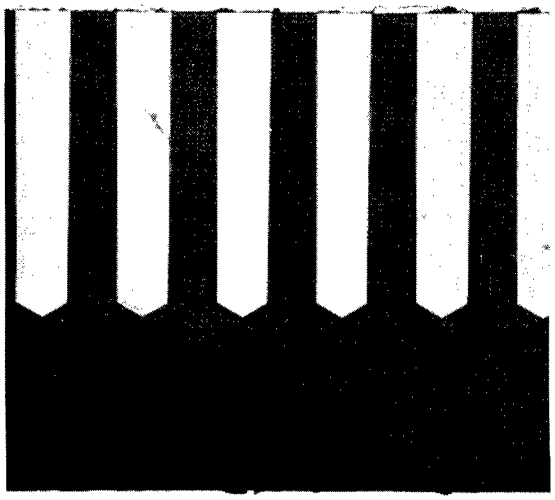


Figure 3-7: Transmission photomicrograph of a microchannel heat sink cross section, anodically bonded to a Pyrex cover plate (top). These channels were formed by anisotropic etching of <110> silicon (100-μm period).

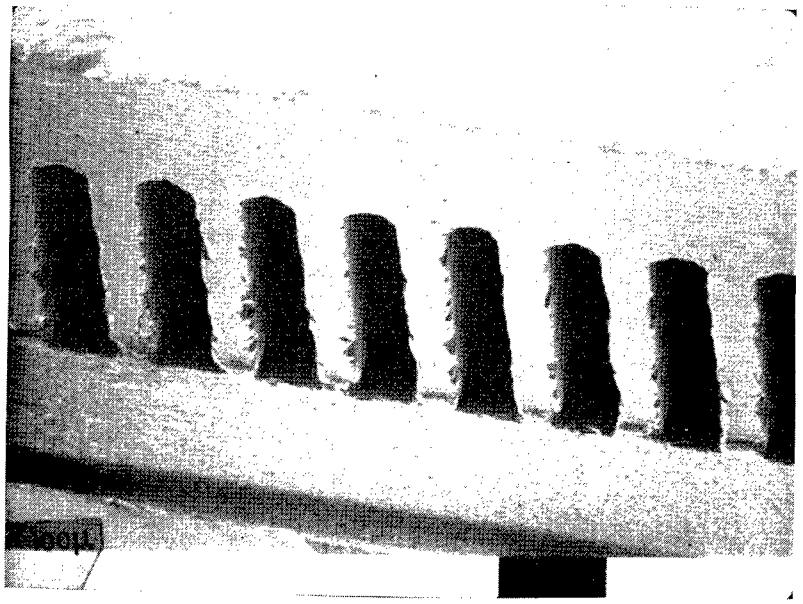


Figure 3-8: SEM of the same microchannel heat sink, viewed at a 45^{o} angle.

3.1.3. Heater Resistor Metallization and Contacts

In order to test the microchannel heat sinks, we required a uniform thin-film heater resistor capable of supplying over 1000 W to a (1 cm)×(1 cm) area on a silicon heat sink. Fig. 3-9 is a sketch of the resistor and contact metallurgy which was devised.

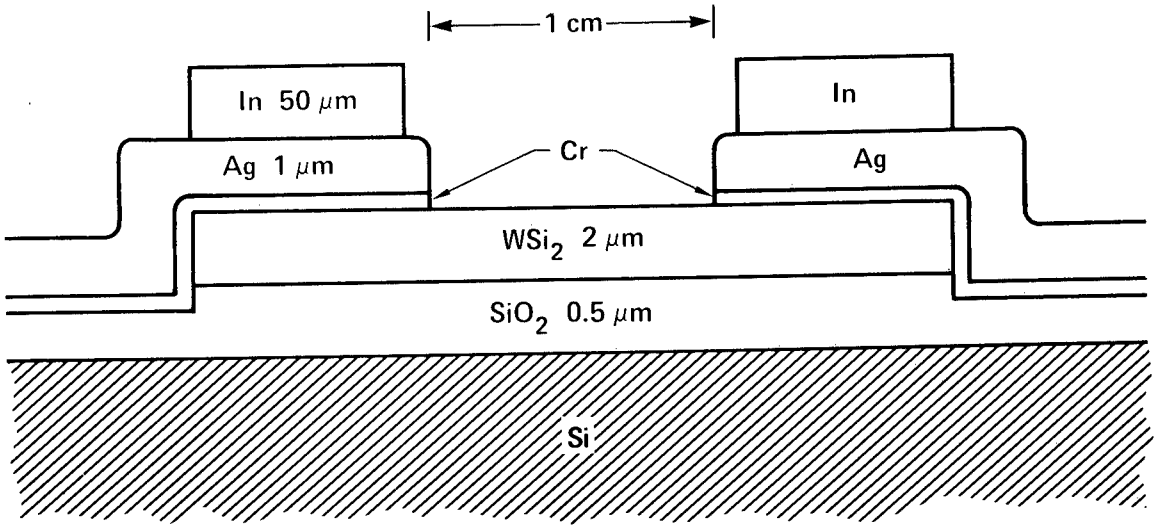


Figure 3-9: Metallization used to fabricate heater resistor and contacts.

The resistor had to be electrically isolated from the underlying silicon so that there is no possibility of current flow (and hence heat generation) in the silicon substrate. On the other hand, an insulating layer would likely have poor thermal conductivity (for amorphous SiO_2, k = 0.014 W/cm-K) [61], so it should not be too thick. We used 0.5 μm of sputtered or thermally grown SiO_2, or sometimes sputtered Pyrex; assuming a conservative dielectric breakdown voltage of 2×10^6 V/cm [62], we conclude that the heater resistor operating voltage should be well under 100 V. These considerations led to a nominal thin-film resistance of 3 Ω/□.

Sputtered tungsten silicide (WSi_2) was chosen as the thin-film resistor, for several reasons. First, its sheet resistance is fairly high (typically 6×10^{-4} Ω-cm, as deposited) [63], hence a 2-μm layer would have the desired resistance. Second, we have found its temperature coefficient of resistance (TCR) is small compared with most metals or with polysilicon. This is undoubtedly due to its amorphous structure, in which the electron mean free path is limited by the short-range disorder and hence is not affected significantly by temperature. Fig. 3-10 shows our measurements of the resistance of a typical sputtered WSi_2 thin-film resistor ($t\simeq0.95$ μm) as a function of temperature. The data indicate a TCR of −.053%/°C (curiously

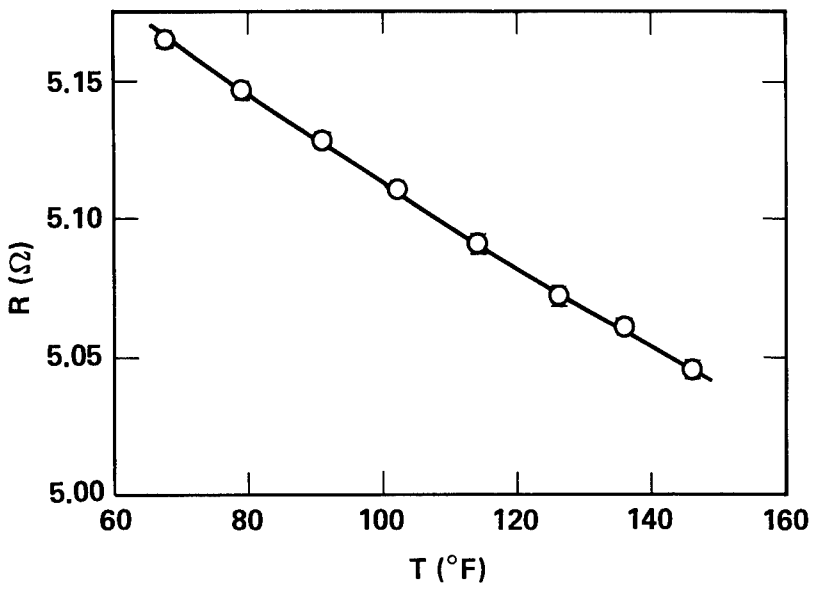


Figure 3-10: Sheet resistance of sputtered WSi_2 (0.95 μm, as deposited) vs. temperature.

this is negative, whereas most metals have a positive TCR); in contrast, the TCR for polysilicon resistors is −2% or −3%/°C! A third reason for using WSi_2 is its chemical inertness; it is not attacked by any of the powerful oxidants which are normally used in cleaning silicon [63].

The WSi_2 was deposited in an rf sputtering system in 20-millitorr argon ambient. The deposition rate was .036 μm/min at a peak voltage of 1.5 kV, and an rf power of 280 W. A 6-inch sputtering target was used to get the best possible uniformity of sheet resistance. Experiments were performed to measure the resistance of a set of adjacent (1 cm)×(1 cm) WSi_2 resistors. From this data we estimate the gradient in sheet resistance (presumably due to a gradient in the deposited film thickness) to be no more than 2.9%/cm. The effects of this slight nonuniformity on our heat-transfer measurements will be discussed in Section 3.2.2.

The WSi_2 was patterned by etching in 1:3:4 HF:HNO_3:HAc, which we found to etch WSi_2 at a rate of 1.8 μm/min. Since photoresist will not withstand this etchant, a photolithographically patterned layer of sputtered SiO_2 was used as the etch mask; the SiO_2 is attacked at the rate of 0.1 μm/min.

Contact was made to the WSi_2 by evaporating a few hundred Angstroms of chromium (for adhesion) at 10^{-7} torr and then, without breaking vacuum, immediately evaporating 1 μm of silver. The room-temperature resistivity of silver is 1.5 $\mu\Omega$-cm, so this resulted in a sheet

resistance of 0.015 Ω/□, which is negligible in comparison with the WSi_2 resistance. The silver was patterned by "lift-off", i.e., evaporation onto a negative photoresist pattern, which was then lifted off by ultrasonic agitation in hot ECOSTRIP®. Gold-clad molybdenum rectangles (20-mil thick) were then soldered to the silver contacts. "Hard" solders such as Au-Ge eutectic were found to cause cracking in the underlying microchannel substrate due to the high soldering temperature and large thermal expansion mismatch. Conventional Pb-Sn solders would often scavenge the thin silver film unless the quantity of solder was carefully limited [64]. Accordingly, we used pure indium foil (2-mils thick; melting point of 156°C) to solder the rectangles; this resulted in no problems with either scavenging or mechanical stresses.

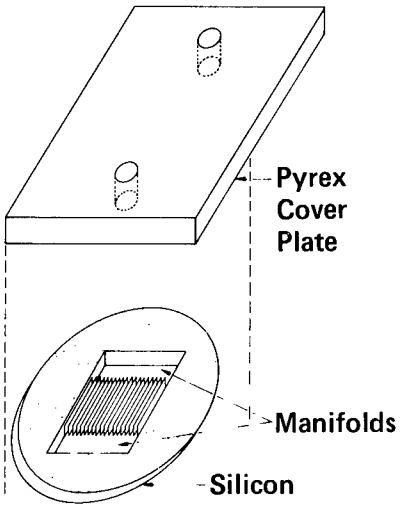
3.1.4. Packaging and Sealing

This section describes the procedures used to header, package, and seal the microchannel heat sinks.

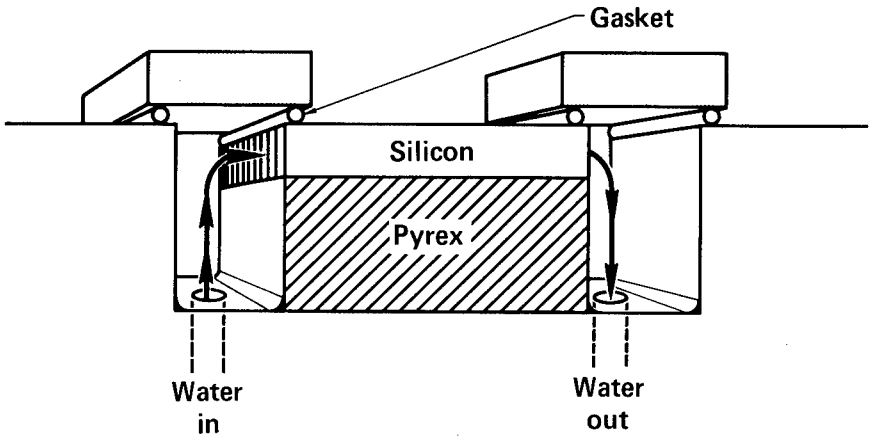
In order to distribute the coolant uniformly among the channels, input and output "headers" or "manifolds" are needed. Figs. 3-11a, 3-11b, and 3-11c show the three approaches which we used. In the former, the headers were etched in the silicon simultaneously with the microchannels. Input and output holes (1.5-mm diameter) were drilled in the Pyrex cover plate using an abrasive slurry. The silicon and Pyrex were then bonded. Several problems with this approach later became evident. First, the KOH etches the unprotected manifolds 40% faster than the deep grooves [50], yet sufficient thickness of silicon must be left under the manifolds so that the silicon will not fracture under pressure (an 80-μm thick, 3-mm wide manifold has a calculated fracture stress of 72 psi). These effects limit the microchannel depths to about 280 μm in a 500-μm thick wafer. Second, the circular holes impart a significant pressure drop (2-3 psi) at the design flow rates of $\sim$10 ml/sec. Third, constructing manifolds in the silicon is difficult when precision sawing is used to manufacture microchannels or pin fins.

Fig. 3-11b illustrates another approach to headering, in which the Pyrex/silicon laminate is diced to expose the microchannels at each end. The heat sink is epoxied in a recessed Lexan® substrate, a gasket material is applied near the edges, and a pair of lids is then applied. The lids are sealed by clamping under a net positive pressure. Coolant may be fed from holes in the recessed substrate. This approach provided the lowest header friction (see Section 3.2.3), but was difficult to fabricate and cumbersome to use.

a)



b)



c)

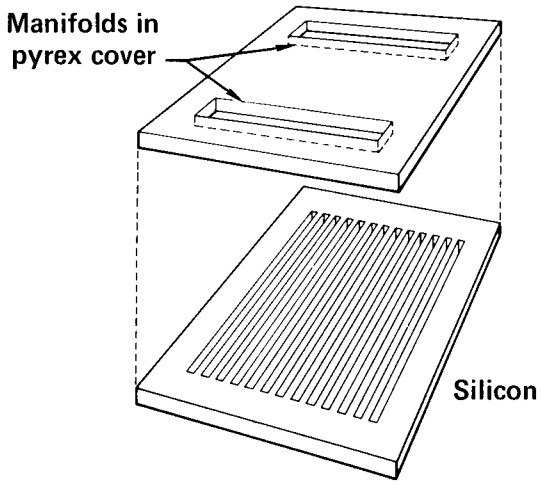


Figure 3-11: Several approaches to packaging and headering silicon heat sinks.

For these reasons, we prefer the configuration of Fig. 3-11c. Here the manifolds are a pair of long rectangular holes in the Pyrex cover plate. The silicon microchannels were fabricated either by ODE or by precision sawing. In the latter case the grooves extend to the ends of the silicon, so after bonding the surfaces, a sealant was needed to prevent leakage at these ends. We applied a controlled amount of epoxy at the ends of the microchannels; capillary action drew the epoxy 1 or 2 mm into the grooves, providing a good seal upon curing. If a perpendicular set of grooves had been cut to make pin-fins, these grooves were similarly sealed. A solder which wets silicon (e.g., Au-Ge eutectic) also worked satisfactorily as a sealant. When the grooves were fabricated by orientation-dependent etching, no sealant was needed because the grooves were designed to terminate within the silicon, as depicted in Fig. 3-11c.

The manifolds were fabricated in 1-mm thick, optically polished Pyrex by sandblasting with a 100-psi jet containing 23-μm Al_2O_3 particles. "Masking tape" was used as a mask for the 3 mm$\times$16 mm manifolds; photogelatin may also be used. It was preferable to blast the side which would not be bonded to the silicon, to ensure that the other side remained optically smooth. The procedure takes about 10 minutes when done with a hand sandblasting tool. The sandblasting process is very directional, producing steep walls (typically 5^{o} from vertical). It is therefore preferable to chemical etching procedures using HF, which would be expected to undercut by at least 1:1 and in fact were found to undercut even more.

The Pyrex was cut to size (25.4$\times$34.5 mm) by precision sawing on a dicing saw with a "resinoid" blade [65]. Metallic-based blades rapidly fail when used on glass, as the diamonds are torn loose from the matrix. Resinoid blades are impregnated throughout with diamonds, and are designed to wear down. A 45-μm diamond grit worked well.

After the Pyrex and silicon were bonded, the Pyrex face was epoxied onto a Lexan® substrate containing matching milled manifolds. These in turn connect to nylon tubing (1/4" O.D. "Parflex"®) epoxied into the Lexan. Reusable connections were made to the assembly by stainless-steel Swagelok® fittings. The epoxy used to seal the manifolds and tubes was Emerson and Cuming semi-rigid Eccobond® 45, a compliant, highly peel-resistant epoxy. This package assembly was very robust and was operated continuously at 50 psi for over a month without problems.

The package was tested with the silicon (resistor) side face up. Temporary, high-conductance electrical contact was made to the gold-clad molybdenum rectangular preforms with phosphor-bronze sheet metal clad in indium foil.

3.1.5. Procedures

The most recent version of the processing schedule for the fabrication of microchannel heat sinks is detailed in Tables 3-1 and 3-2. Table 3-1 lists the overall fabrication sequence. Table 3-2 gives details of some subprocedures referred to in Table 3-1. Nondestructive measurement of groove depth is performed using an optical microscope with a calibrated stage height and by focusing alternately at the top and bottom of the microchannels. It is important to make sure that all trapped water (tenaciously held due to the strong capillary attraction) has been removed from the grooves, in order to prevent erroneously low depth measurements due to refraction. This cannot be achieved merely by blowing with an N_2 jet, so we always heated the thoroughly-rinsed substrate to 100°C before measuring channel depth. The accuracy of this measurement is about ± 5 μm. The microchannels must also be baked prior to immersing in ECOSTRIP®; otherwise the trapped water reacts with the ECOSTRIP® to yield sulfuric acid which corrodes the heater contact metallization.

It is difficult to thoroughly rinse the microchannels using normal techniques, due to the relatively stagnant liquid; ultrasonic cleaning may work but sometimes fractured the substrates. To solve this, the substrate is first boiled in a solvent series (TCE, acetone, methanol); the grooves provide good nucleation sites and hence good agitation which removes particulates. For the final rinse, the substrate is spun at 7000 RPM on a chuck while being squirted with solvents. The large centrifugal forces produce radial pressure drops sufficient to thoroughly flush contaminants from the channels as they are being rinsed.

Table 3-1: Fabrication schedule for silicon microscopic heat sinks

Step	Process
1. Starting material	2", <110>, 500-μm thick, preferably double-polished, lightly-doped (>1 Ω-cm) Measure flatness with optical flat Label flattest side as "back"
2. Oxidation	**RCA clean** 1100°C, 5' dry, 35' wet, 5' dry (0.5 μm nom.), back side up
3. Splay pattern lithography	**NPR photolith** (back side) Front-surface protect Buffered HF etch, DI rinse 5', Nanospec inspect **DEWAX, STRIP** Inspect SiO_2 for pinholes or scratches
4. Splay pattern etch	50-50 KOH @ 70°C for 2 hrs (nom. depth 94 μm) Long DI rinse Identify <111> plane angle Strip SiO_2 in 10:1 HF
5. Oxidation	1200°C, 5' dry, 90' wet, 5' dry (back side up) slow push/pull! (t_{ox} = 1.1 μm nom.)
6. Microchannel lithography	**NPR photolith** (back), aligned w/ best <111> plane Front-surface protect Buffered HF etch, DI rinse 5', Nanospec inspect **DEWAX, STRIP**, inspect front SiO_2 for pinholes
7. Microchannel etch	Fresh, clean, 50-50 KOH 52°C (tilted) Etch 34 hrs (check @27 hrs); nom. depth 400 μm Long DI rinse
8. Prepare for anodic bonding	Saw to 34.5 mm × 34.5 mm Ultrasonic solvent-clean **RCA clean** w/10:1 HF strip after H_2O_2/H_2SO_4. (agitate during SiO_2 strip)
9. Cover plate fabrication	Start w/optically polished Pyrex 0.8 mm thick Saw to (34.5 mm) × (34.5 mm) (resinoid blade) Cut headers (sandblast or ultrasonic drill) Ultrasonic solvent-clean H_2O_2/H_2SO_4 clean
10. Cover plate bond	Anodic bond @ 400°C Measure laminate curvature with optical flat Saw to final size (25.4 mm × 34.5 mm)

continued

Step	Process
11. Heater resistor fabrication	Sputter SiO_2 (0.5 μm) on front for isolation Sputter WSi_2, 65' (2.34 μm nom.) Sputter SiO_2 (0.2 μm) for etch mask
12. Resistor lithography	**NPR photolith** (1 cm $\times$ 2 cm rectangle) Back-surface protect (seal headers) Buffered HF etch **DEWAX, STRIP** Etch WSi_2 in 1:3:4 HF:HNO_3:HAc ($\sim$75 sec), continue until SiO_2 mask gone ($\sim$60 sec more)
13. Contact metallization	**NPR photolith** (contact rectangles) Evaporate Cr/Ag (1 μm) Lift-off in hot, dry ECOSTRIP (ultrasonic agitation) Solvent clean (TCE/Ace/Meth)
14. Bond contacts	Solder Au-clad Mo rectangles using indium foil
15. Package	Epoxy to Lexan® containing milled manifolds

Table 3-2: Standard fabrication subprocedures

Subprocedure	Process
NPR photolith	Spin @5000 RPM for 30 sec 20' prebake @90°C 6.0" exposure (clean mask) 15" develop, 20" rinse, inspect 15' postbake @120°C O_2 plasma descum 60", 500 Watts, 100 SCCM
DEWAX	Load into Teflon boat TCE boil, slide wafer off TCE boil
STRIP	Transfer to clean beaker (no Teflon boats) H_2SO_4/Chromic acid strip 1:1 H_2SO_4/H_2O_2 5:1:1 $H_2O/NH_4OH/H_2O_2$ (omit with Pyrex) 5:1:1 $H_2O/HCl/H_2O_2$ (only if going into furnace) 50:1 HF dip DI rinse, N_2 dry
RCA clean	Same as **STRIP**, except no H_2SO_4/Chromic strip

3.2. Experiments

3.2.1. Test Apparatus and Techniques

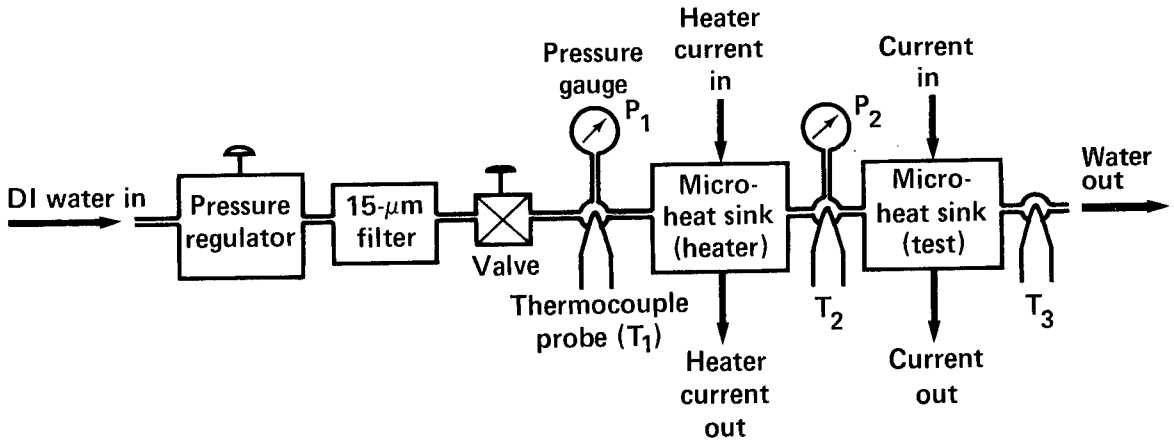


Figure 3-12: Apparatus for heat sink flow-friction and heat-transfer measurements.

Fig. 3-12 diagrams the experimental apparatus used to test the silicon heat sinks. Stainless-steel Swagelok® fittings were used whenever possible, to minimize corrosion. Deionized water was supplied to a precision pressure regulator (Moore Nullmatic® 40E-50). This regulator has two stages of internal mechanical gain (i.e., two internal diaphragms), and hence was far more stable than standard single-diaphragm regulators. The typical supply sensitivity was $\partial P_{out}/\partial P_{in} = 1/150$. Although designed for handling air, the regulator performed well with 18-MΩ deionized ("DI") water for about a year before needing replacement due to pitting of the brass pilot-valve seat. The DI water flowed through a high-flow ($C_v = 0.21$ gpm/psi$^{1/2}$) 15-μm filter (Nupro® SS-4F-15), a valve, and then into either a single silicon heat sink or a series combination of two silicon heat sinks which were clamped to a Lexan substrate. The latter arrangement enables us to use the upstream heat sink as a heater to preheat the water which then supplied the second (test) heat sink. The water pressure was probed at the input to both heat sinks (P_1 and P_2) using high-precision Ametek® test gauges (1/4% full-scale accuracy; 15, 30 or 60 psi full scale). The gauges were carefully adjusted so that all indicated zero hydrostatic gauge pressure when the water was not flowing. Great care was taken to insure that the gauges and connecting lines were devoid of large air bubbles, because the expansion and contraction of such bubbles with pressure changes would cause variations in the hydrostatic pressure (0.43 psi/ft of water).

dc power was supplied to the WSi_2 resistor contacts on each heat sink from a pair of

regulated supplies. In some experiments a high-power programmable operational amplifier (Kepco BOP 50-8M) was used to supply transient pulses of power. Supply voltages were measured directly at the resistor contact point with a digital voltmeter. Supply currents were determined from the voltage across a precision 0.01-Ω metal film resistor (±0.25% accuracy) in series with the supply.

The heater resistor surface temperature was normally probed using a movable copper-constantan beaded-junction thermocouple. (In a few cases where a fixed-location temperature measurement was desired, the thermocouple was bonded to the resistor surface with a thermally conductive epoxy.) As shown in Fig. 3-13, 5-mil thermocouple wires were threaded through 1/64" holes in a mullite insulator.

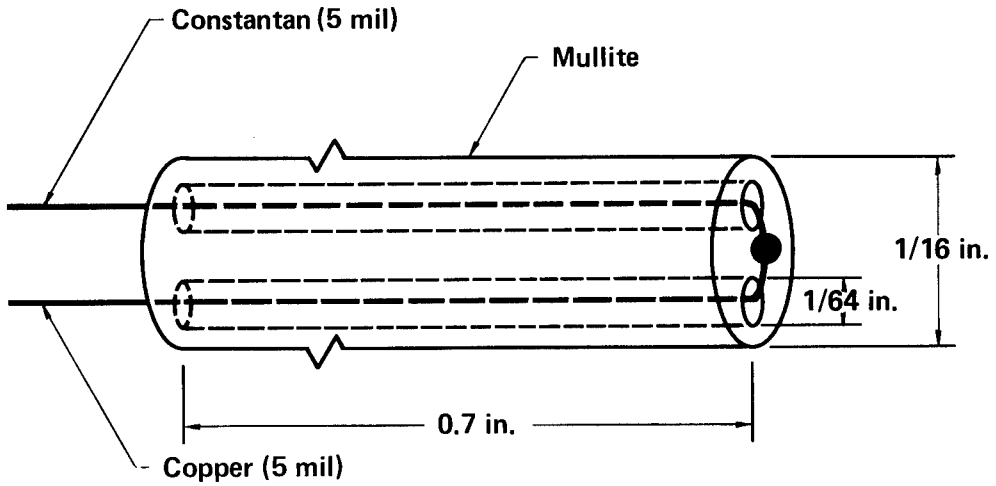


Figure 3-13: Thermocouple probe to measure surface temperature. The bead is epoxied flat against the end of the mullite insulator.

The junction bead was bent flat against the end of the mullite and a very small amount of low-thermal conductivity epoxy was used to hold the bead in place, taking care not to cover up the exposed portions of the bead which would contact the resistor. The probe was mounted on a 3-axis micromanipulator and lowered into contact at any desired location on the heat sink surface. A heat-sinking compound (Omegatherm® 201) was used to reduce thermal contact resistance. Probe contact pressure was adjustable (the heat sinks rested on a compliant substrate).

Fluid temperatures were also measured by bare copper-constantan beaded-junction thermocouples. Probes were located at the input and output of each heat sink device and at the drain point. The thermocouple wires were threaded through ceramic insulators, sealed

with epoxy, and installed in Swagelok® fittings so that the thermocouple bead was positioned in the center of the pipe flow.

3.2.2. Data Analysis and Experimental Errors

Three types of measurements were involved in this work: electrical, flow-friction, and thermal. Of these, the thermal measurements present the greatest possibility of error, and we shall discuss them in the most detail.

The copper-constantan thermocouple temperatures were measured using an Omega 2176A 10-channel digital thermometer having a resolution of 0.2°F. Due to material variations, the specifications state that the thermocouples may have an offset error of up to ± 1°C. However we are usually interested in temperature **differences** rather than in absolute temperatures, so the offset error can be determined and subtracted out in those cases. By choosing all thermocouples from the same lot, we verified that all thermocouples agreed within 0.2°F of each other over the entire temperature range of interest (20°C-120°C).

The measurement of temperature in a flowing fluid introduces possibilities for error. We wish to measure the "mixed mean temperature" T_c (also known as the "cup-mixing temperature"). A thermocouple probe, however, measures temperature at a single point. If the fluid has just exited from a heat sink, the temperature profile may not be spatially uniform across the pipe cross section, hence the measured temperature could deviate significantly from T_c. On the other hand, placing the thermocouple far downstream, where the fluid is presumably thermally mixed, can lead to errors if a significant heat loss to the ambient occurs in the long pipe. To investigate these considerations, an experiment was performed in which thermocouple probes having identical characteristics (so far as could be measured with the digital thermometer) were located at four points in the output flow: (1) within the output manifold of an operating water-cooled silicon heat sink; (2) just outside the output manifold (within the Swagelok fitting); (3) 50 cm downstream of the heat sink; and (4) 100 cm downstream. The temperature drop in the second 50 cm ($T_4 - T_3$) was found to be small and equal to the drop in the first 50 cm ($T_3 - T_2$), and both were proportional to $T_3 - T_A$, where T_A is the ambient air temperature. This suggests that the temperature profile of the water is already well mixed by the time it reaches the output fitting (position 2), and that $T_3 - T_2$ and $T_4 - T_3$ are both due to heat loss through the tubing (Parflex® clear PVC tubing, .250" OD, .170" ID). The measured heat leak per unit length of tubing was 0.76 W/°C-m. If we assume the tubing to be a good conductor of heat, this implies a tube-to-air convective heat-transfer

coefficient of 3.81×10^{-3} W/cm^2-K, hence a Nusselt number of 9.1. Since Nu$\simeq$0.615 Re$^{.466}$ for forced convection around a cylinder in this regime [66], an air velocity of 80 cm/sec (157 fpm) would explain the heat loss. Since the laminar-flow hood under which the experiments were conducted has a nominal air velocity of 150 fpm, this seems to be a correct explanation of the heat loss.

We thus conclude that measuring the water temperature at the output fitting (T_2) gives an accurate value for the mixed mean temperature at that location. This has been verified over the full flow range used in our experiments. We found that the thermocouple installed directly within the output manifold (T_1) gave erratic results, presumably because the thermal mixing was not yet complete there. The possible heat leak between the sites of T_1 and T_2 is easily calculated to be negligible ($\ll 0.2^{\circ}$F); hence T_2 was considered to be an accurate and adequate measurement of the output mixed mean water temperature.

The thermocouple probe of Fig. 3-13, used for temperature measurements of the resistor surface, was another source of potential error due to heat leaks to the ambient. One desires that the thermal contact resistance of the bead to the surface be much lower than the thermal resistance of the bead to the ambient. The thermal resistance θ of a spherical bead in contact with a surface is known [13] to be $\theta_{probe} = 1/2\pi\rho k_m \cdot \ln(Y_{max}/Y_{min})$, where ρ is the bead radius ($\rho = .019$ cm), Y_{min} is the surface roughness (estimated to be 1.43 μm), Y_{max} is the largest contact gap ($\simeq$125 μm), and k_m is the thermal conductivity of the interfacial material. If the interfacial material is air, we calculate $\theta_{probe} = 7.2^{\circ}$C/mW. If the interfacial material is Omegatherm® 201 (a thermally conductive paste having $k = .023$ W/cm-K), then $\theta_{probe} = 0.08^{\circ}$C/mW.

The heat leak to ambient through the leads can be calculated using standard formulas [67]: $\theta_{leak} \simeq (hck_{lead}A_{lead})^{-1/2}$ where c is the mullite circumference ($c = (2\pi R) = 0.5$ cm), k_{lead} is the lead thermal conductivity ($k_{lead} \simeq 3$ W/cm-K), A_{lead} is the lead area ($A_{lead} = 1.267 \times 10^{-4}$ cm^2), and h is the heat transfer coefficient from the leads to to ambient referenced to the outer surface area of the mullite. Now $h^{-1} = h_{lead\text{-}mullite}^{-1} + h_{mullite}^{-1} + h_{mullite\text{-}ambient}^{-1}$. Simple geometrical approximations give $h_{mullite} \simeq 0.28$ W/cm^2-K. From the previous discussion, we estimate $h_{mullite\text{-}ambient} = 0.008$ W/cm^2-K. Finally we assume a stagnant air layer between the leads ($r_i = 0.127$ mm) and the mullite holes ($r_o = 0.397$ mm), thus

$$h_{lead\text{-}mullite} \simeq 2k_{air}/[R \ln (r_o/r_i)] = 0.006 \text{ W/cm}^2\text{-K}.$$

Thus $h \simeq 3.3$ mW/cm^2-K, whence $\theta_{leak} \simeq 1.24^{\circ}$C/mW in this simple model. Furthermore, the

bulk of the heat is lost over a distance $(kA/ch)^{1/2} = 0.47$ cm, so there is no point in making the mullite insulator longer than its actual height of 1.8 cm.

If we use the thermally conductive paste, we predict a heat leak of $\theta_{probe}/\theta_{leak} = 0.08/1.24 = 6.4\%$, which is a significant error, but small enough that it could be corrected for. If the thermally conductive paste were not used, the heat leak would be enormous ($\theta_{probe}/\theta_{leak} = 7.2/1.24 \gg 1$).

The **actual** heat leak was measured by passing preheated water (from heat sink 1) through heat sink 2, and measuring the the surface temperature of heat sink 2 with the thermocouple probe. Due to the very high heat-transfer coefficient of the heat sink compared with that of the ambient air, the surface temperature is equal to the water temperature within 0.5%. Fig. 3-14 is a plot of the indicated probe temperature as a function of the known surface temperature ($T_A = 67^oF$ in this experiment). A least-squares fit to this curve gives a slope of $(T_{probe} - T_A)/(T_{surface} - T_A) = 0.925$; this corresponds to an actual heat leak of $\theta_{probe}/\theta_{leak} = 8.1\%$, which is quite close to the predicted error of 6.4%. This number was very reproducible (less than 10% variation), hence all probe measurements were corrected for this heat leak (T_A was always noted during an experiment). Since the correction is 8%, and reproducibility is better than 10% of that, we conclude that our corrected temperature measurements have an accuracy of $\pm 1\%$.

Electrical measurements of dc supply voltage and current were made with digital meters which were verified to have 0.1% accuracy. Water flow measurements were made by weighing the amount of water expelled in a given time interval (rather than by measuring volume), thus no error is introduced by water density variations. Weight measurements were accurate to within $\pm 0.1\%$, and elapsed time measurements were determined by repeated experiments to be within ± 0.2 sec (3σ error).

In our data analyses, it was assumed that the electrical power input to the heater resistor was, in the steady state, all transferred to the water, to within the accuracy of our measurements. In most heat exchanger experiments, this would not be a good assumption. However, the heat-transfer coefficient of a typical micro-heat sink surface is very high (typically >10 W/cm^2-K). In contrast, typical convective heat-transfer coefficients for 80-cm/sec air over planar surfaces are typically 3 orders of magnitude lower. This assumption was experimentally verified (and the calibration of our equipment confirmed) by an energy balance test:

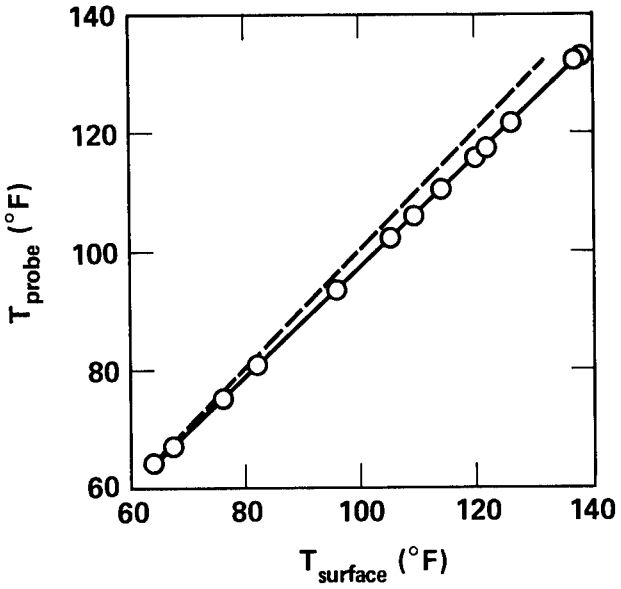


Figure 3-14: Thermocouple probe calibration ($T_A = 67^oF$). The dashed line ($T_{probe} = T_{surface}$) would be for a perfectly thermally insulated probe.

Input current	$I = 8.65 \pm .05$ A
Input voltage	$V = 30.2 \pm .05$ V
Mass flow	$\dot{w} = 3.060 \pm .016$ gm/sec
$T_{input} = 62.6 \pm 0.2^oF$,	$P_{input} = 16.8$ psig $\Rightarrow h_{in} = 71.6 \pm 0.5$ J/gm
$T_{output} = 99.2 \pm 0.2^oF$,	$P_{output} = 0.1$ psig $\Rightarrow h_{out} = 156.4 \pm 0.5$ J/gm

Thus for an input power of 85.4 ± 0.7 J/gm, the water exhibited an enthalpy change of

$$\Delta h = h_{out} - h_{in} = 84.8 \pm 0.7 \text{ J/gm},$$

which means that the energy balance is confirmed to within the measurement error. At much higher water temperatures ($T \simeq 90^oC$), small deviations in the energy balance were observed (Δh was low by 2 or 3 J/gm). We attribute this discrepancy to the heat of solution of dissolved gases which come out of solution when the water is heated; the evolved gas bubbles are clearly visible in the output tube.

In Section 3.1.3 we noted a slight (up to 2.9%/cm) spatial gradient in heater sheet resistance due to nonuniformities of the sputtering system. While this is too small to greatly alter the temperature profiles, it will be useful to know its effect on local heat flux when making accurate comparisons with experiment. Fig. 3-15 is a sketch of the heater

configuration, in which a square of side L = 2S is fed current by two equipotential contacts at opposite edges.

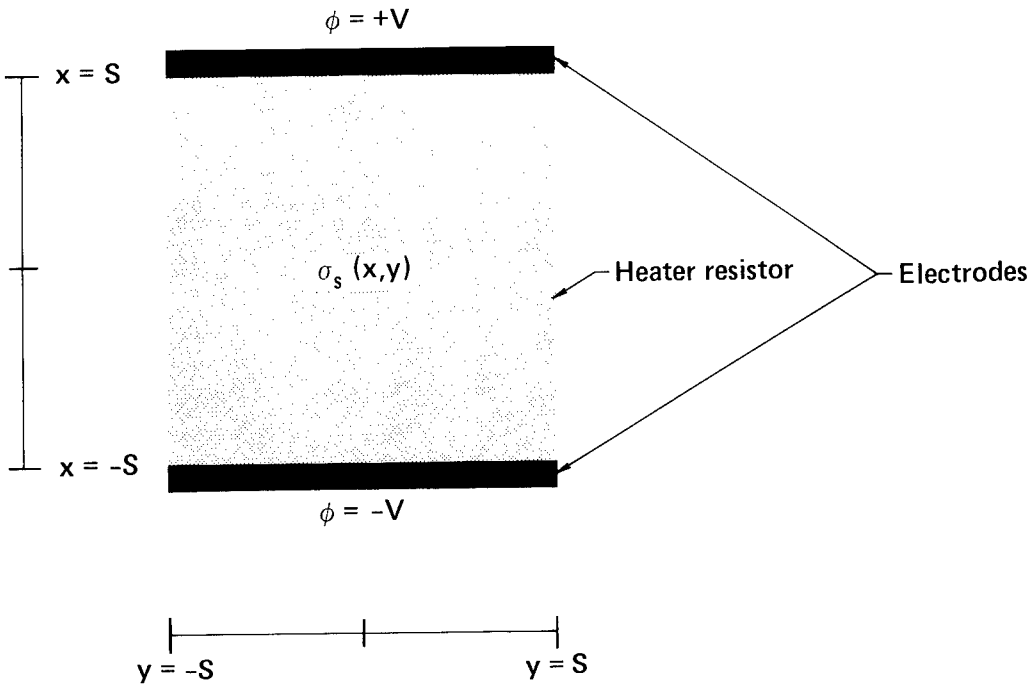


Figure 3-15: Notation used to calculate effects of nonuniform heater sheet resistance.

Assuming a linear gradient in sheet conductance σ_s, we write

$$\sigma_s(x,y) = \sigma_o(1 + \kappa x/S + \lambda y/S),$$

where $\kappa \ll 1$ and $\lambda \ll 1$ are the normalized x and y components of $(\nabla\sigma_s)$. Let φ be the potential; the boundary conditions are then $\varphi = +V$ at $x = S$ and $\varphi = -V$ at $x = -S$. From Ohm's law, the sheet current is

$$J_s = -\sigma_s \nabla\varphi .$$

From continuity,

$$-\nabla \cdot J_s = \nabla \cdot (\sigma_s \nabla\varphi) = [\sigma_s \nabla^2\varphi + (\nabla\sigma_s)\cdot(\nabla\varphi)] = 0 . \qquad (3.1)$$

Let $\varphi(x,y) = \varphi_o + \varphi_1$ where $\varphi_o = xV/S$ is the "unperturbed" solution, i.e., for the case of uniform sheet conductivity, in which $\sigma_s(x,y) = \sigma_o$. Then $\nabla^2\varphi_o = 0$, hence from Eq. (3.1),

$$\sigma_s \nabla^2\varphi_1 = -(\nabla\sigma_s)\cdot\nabla(\varphi_o + \varphi_1) \simeq -(\nabla\sigma_s)\cdot\nabla\varphi_o = -(\partial\sigma/\partial x)(V/S) = -\sigma_o \kappa V/S^2.$$

Thus

$$\nabla^2\varphi_1 \simeq (-\kappa V/S^2)/(1 + \kappa x/S + \lambda y/S) \simeq -\kappa V/S^2$$

to first order. This may be solved using the boundary condition that $\varphi_1 = 0$ at $x = \pm S$, to yield $\varphi_1 \simeq \kappa[1 - (x/S)^2]V/2$. The local power density is

$$\dot{q}''(x,y) = \sigma_s|\nabla\varphi|^2 \simeq \sigma_o(1+\kappa x/S+\lambda y/S)\cdot[V/S - x\kappa V/S^2]^2 \simeq$$

$$\simeq \sigma_o(V/S)^2\cdot[1-\kappa x/S+\lambda y/S],$$

to first order in κ and λ. The total heat generated $\dot{q}$ is thus determined by the average sheet conductivity σ_o, i.e., $\dot{q} = \sigma_o V^2$. Furthermore, the **local** heat flux at the geometric center of the heater ($x = y = 0$) is equal to the **average** heat flux: $\dot{q}''(0,0) = \dot{q}/S^2 = \sigma_o V^2/S^2$. Thus we do not have to correct for this conductivity gradient when performing heat-transfer measurements, provided such measurements are made at the center of the thin-film heater square.

3.2.3. Flow-Friction Measurements

Experiments were performed to verify that the flow-friction characteristics of silicon micro-heat sinks were as predicted from theory. Two types of headering schemes were tested: the direct (end-fed) header shown in Fig. 3-11b, and the side-fed header shown in Fig. 3-11c, in which the flow must turn a corner as it enters the channels. As discussed in Section 2.2.3, the total pressure drop including headers and entrance effects should take the form given by Eq. (2.44):

$$P = P_{core}(1+K\chi/4\Phi),$$

where $P_{core} = 2\mu\Phi Lv/D^2$, $\chi \equiv D\cdot Re/L = vD^2\rho/\mu L$, and K is a loss factor due to entrance and exit effects. Another way to write this would be

$$P = 2\Phi_\chi \mu Lv/D^2,$$

where $\Phi_\chi = \Phi + K\chi/4$. Thus a plot of Φ_χ against χ would yield a straight line, with its intercept at $\chi = 0$ giving the fully-developed laminar-flow friction factor $\Phi = c_f Re$ (predicted in Fig. 2-5), and its slope giving the loss factor K associated with entrance/exit effects. Fig. 3-16 shows typical experimental results for plate-fin silicon micro-heat sinks.

The experimental apparatus was as shown in Fig. 3-12; sometimes the water was preheated and sometimes not. The measured quantities were: the water mass flow Δw in a given time interval Δt, the water temperature T at each end of the heat sink (virtually identical, since no electrical power was being supplied to the sample), and the pressure drop P across the heat sink. In addition, the heat-sink geometrical parameters L, w_c, H and n (the number of microchannels) were measured using optical and electron microscopy either before the experiment (during fabrication) or afterwards (by sectioning the sample and examining it). In most cases the measurement errors were estimated to be:

Δw ($\pm$0.1 gm);

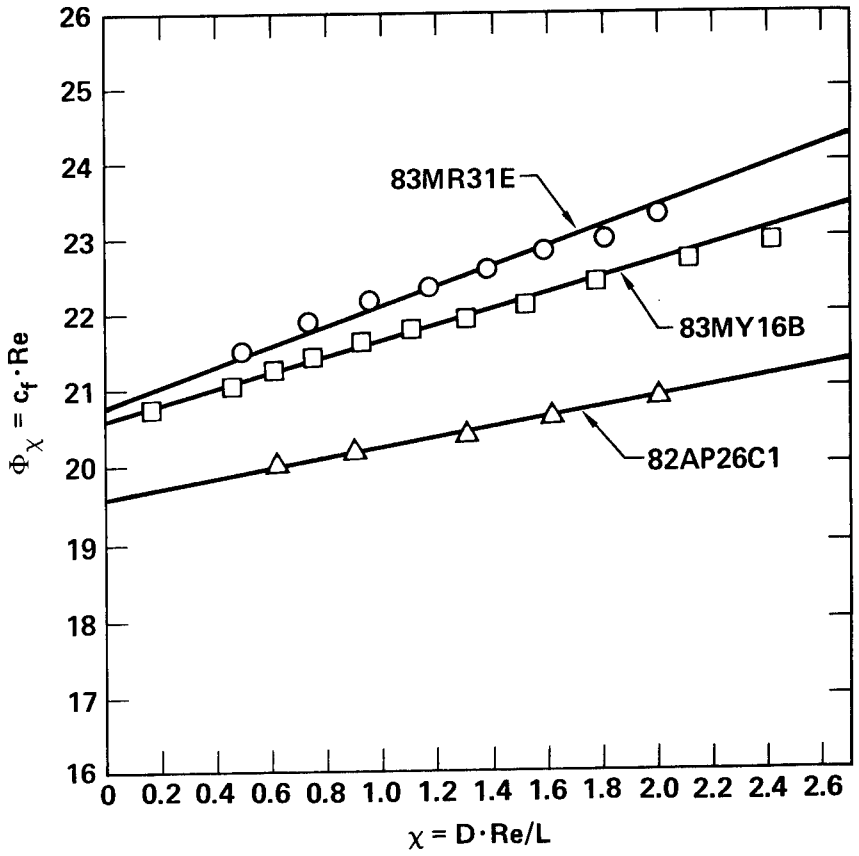


Figure 3-16: Flow friction parameter $\Phi_\chi \equiv c_f Re$ as a function of $\chi \equiv D \cdot Re/L$ for plate-fin microchannel heat sinks (includes header losses).

Δt ($\pm$0.2 sec);

T ($\pm$1.0°C; this error is a possible systematic offset in the thermocouple);

P ($\pm$0.1 psi);

L ($\pm$0.05 cm for side-fed headers; $\pm$0.01 cm for end-fed headers);

w_c ($\pm$1.0 μm);

H ($\pm$5 μm);

n (no error; it is an integer).

χ is calculated from

$$\chi = \Delta w D^2 / n \mu L w_c H(\Delta t) .$$

Φ_χ is calculated from

$$\Phi_\chi = nD^2 w_c H \rho P(\Delta t)/2\mu L(\Delta w),$$

where $D = 2w_c H/(w_c + H)$ is the hydraulic diameter of each microchannel. The values of μ and ρ are appropriate to the measured water temperature. Since the channels are not perfectly rectangular, w_c and H represent average measured values of channel width and depth, respectively.

Inspection of Fig. 3-16 shows that Φ_χ does indeed exhibit a linear dependence on χ, as predicted. The slope (K/4) and intercept Φ were determined by a least-squares fit and are tabulated in Table 3-3.

Sample	82AP26C1	83MR31E	83MY16B
Type Header	End-fed	Side-fed	Side-fed
Fabrication	Sawn	Etched	Etched
Channel Dimensions:			
L (cm)	2.00±.01	1.55±.05	1.48±.05
w_c (μm)	64.2±1.0	54.3±0.7	59.3±1.0
$w_c + w_w$ (μm)	200	100	100
H (μm)	284±3	351±6	376±6
Maximum Re	900	330	390
K (expt.)	2.6	5.3	4.2
Φ (expt.)	19.6±1.0	20.8±1.1	20.6±1.1
Φ (theory)	18.6	20.0	19.9

Table 3-3: Summary of flow-friction properties of plate-fin silicon microchannel heat sinks.

The experimental error in determining Φ is typically 5%; most of this error is due to the uncertainty in measuring channel width w_c. Φ is in agreement with the predictions of Fig. 2-5 within this 5% experimental error, regardless of whether the channels were etched or sawn. Thus there are no evident problems with fouling or clogged channels. However, the entrance/exit loss factor K for the end-fed headers (shown in Fig. 3-11b) is larger than would be expected from the simple model of abrupt expansion and contraction: $K \simeq 3.0$ instead of 1.6. The side-fed headers (Fig. 3-11c) are even lossier: $K \simeq 4$ or 5. In the case of side-fed headers the discrepancy is not surprising, in view of the fact that the flow "turns a corner". In turbulent pipe flow, a rounded 90° elbow can result in friction loss factors $\Delta K = 0.4$ to 0.9 for each bend; a square 90° contributes $\Delta K = 1.3$ to 1.9 to each bend. In the case of end-fed

headers, the discrepancy is harder to explain; in any case, the microscopic geometry of the headers in their package can neither be readily examined nor perfectly controlled. At the recommended operating points, (calculated in Section 2.2.3), these discrepancies in flow friction are not sufficiently serious to warrant further investigation into their causes. The increase in thermal resistance is less than 10% for an optimized design, as shown in Section 2.2.3.

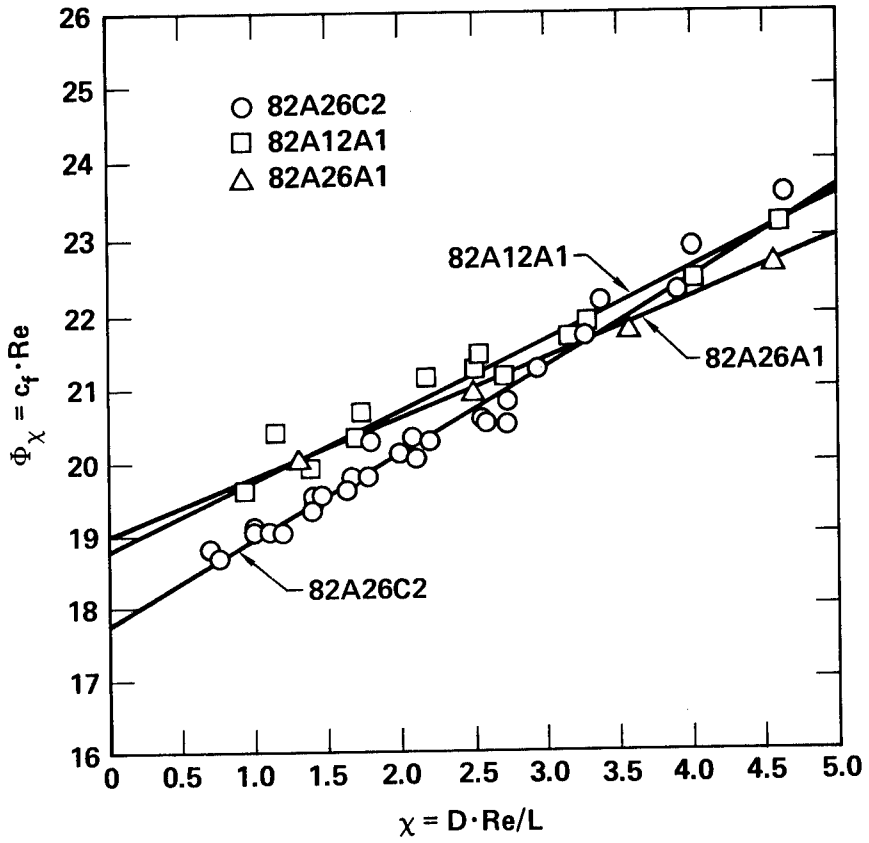


Figure 3-17: Flow-friction parameter $\Phi_\chi \equiv c_f Re$ as a function of $\chi \equiv D \cdot Re/L$ for pin-fin microchannel heat sinks.

Rectangular **pin-fin** structures (fabricated by precision sawing) were also examined. Fig. 3-17 plots Φ_χ vs. χ for such structures; all had end-fed headers (Fig. 3-11b). Here we have calculated Φ_χ and χ by ignoring the fact that the fins are interrupted, i.e., by treating the structures as plate fins. While Kays and London [17] use a different definition of Reynolds number and friction factor for interrupted structures, our purpose here is to show the correlation with the plate-fin structures, rather than to achieve consistency in presentation. Table 3-4 tabulates Φ and K. The main point to be noted here is that the results are rather

Sample	82AP12A1	82AP26A1	82AP26C2
Type Header	End-fed	End-fed	End-fed
Fabrication	Sawn	Sawn	Sawn
Channel Dimensions:			
L (cm)	2.00±.01	2.00±.01	2.00±.01
w_c (μm)	102.5±1.0	88.9±1.0	64.2±1.0
$w_c + w_w$ (μm)	160	200	200
Interruption width/period (μm)	56/160	60/200	60/200
H (μm)	367±6	255±6	284±6
Maximum tested Re	900	700	760
K (expt.)	3.6	3.4	4.7
Φ (expt.)	18.8±1.0	19.0±1.0	17.8±1.0
Φ (theory)	17.8 *	16.9 *	18.6 *

* Calculated as if no interruptions existed

Table 3-4: Summary of flow-friction properties of pin-fin silicon microchannel heat sinks.

similar to what we'd expect for uninterrupted plate fins, at the Reynolds numbers which we investigated.

To summarize, the experimental flow-friction characteristics of microscopic silicon plate-fin heat sinks are as predicted by theory, except that the header losses appear to be somewhat larger than expected. The discrepancy is the order of one or two times the velocity head for each of the two headers, which is certainly a plausible discrepancy in view of the uncertainties about the flow in the headers. Pin-fin heat sinks seem to differ only slightly at the Reynolds numbers tested (up to 900), implying that the fin interruptions do not modify the flow significantly. Undoubtedly at higher Reynolds numbers the differences would become evident, but this is outside our design region.

3.2.4. Heat-Transfer Measurements

A series of experiments was performed to investigate the heat-transfer characteristics of silicon micro-heat sinks. Both plate-fin and pin-fin structures were tested. A uniform heat flux was generated in a rectangular thin-film WSi_2 resistor of nominal dimensions $L \times W = (1\text{ cm}) \times (1\text{ cm})$. The cooling channels covered a substantially larger rectangular area so as to make the registration between the heater resistor and the cooled substrate noncritical, as well as to make the packaging more convenient. The channel array (cold plate) dimensions are denoted by the subscript "s", i.e., length L_s and width W_s. This discrepancy between the cold-plate area and the smaller heated area implies that not all the input flow is effective in cooling. For example, if $W_s = 1.5$ cm whereas $W = 1.0$ cm, then approximately 2/3 of the coolant flows underneath the heater resistor. We say "approximately" because when the resistor is generating heat, the water temperature underneath the heater will increase, reducing its viscosity, hence slightly more than 2/3 of the total coolant flows underneath the heater. For plate-fin micro-heat sinks (uninterrupted channels) this effect may be estimated by measuring the total flow rate f_{cold} when the heater is off and the flow rate f_{hot} when the heater is on. Then we estimate the coolant flow rate under the heated area to be $f = f_{hot} - (1 - W/W_s)f_{cold}$, i.e., we assume the flow in the unheated area to be unchanged. For pin-fin structures the problem is more complicated because the coolant is not confined to discrete channels, and hence is free to cluster at the areas of higher temperature (lower viscosity). In fact, this clustering effect is the main experimental benefit which we have observed from pin-fin structures. Lateral heat spreading is an additional complication caused by the relatively large cold plate; significant effects were observed within 2 mm of the resistor edge, as predicted in Section 2.2.5.

The first set of experiments involved simply measuring temperature T at the 90% downstream location ($x = 9$ mm) for progressively higher powers. The thermal resistance at maximum tested power is tabulated in Table 3-5. Note that the use of a substrate length L_s which is longer than the 1-cm channel width ($L_s \geq 1.4$ cm in our samples) has handicapped us somewhat. Since optimized thermal resistance scales as $L_s^{1/2}$ for constant pressure, we could extrapolate the performance of sample 81F9 to $R = .077$ cm^2-°C/W and that of sample 81D7C3 to $R = .067$ cm^2-°C/W for the case where $L_s = 1.0$ cm and $P = 30$ psi; this requires scaling down all dimensions by $(L/L_s)^{1/2}$. These numbers are close to those predicted in the elementary optimization procedure of Section 2.1 for $L = 1$-cm heat sinks. The pin-fin heat sink has slightly better thermal performance than the plate-fin heat sinks, not because of any turbulence effects (as explained in Section 3.2.3), but simply because the viscosity of the

Sample	82A26C1	80D6	80D19	81F9	81D7C3
Type Header	End-fed	End-fed	End-fed	End-fed	End-fed
Fabrication	Etched	Etched	Etched	Etched	Sawn
Channel Dimensions:					
L_s (cm)	2.00	1.40	1.40	1.40	2.00
W_s (cm)	1.5	2.0	2.0	2.0	1.5
w_c (μm)	64	56	55	50	55 (pins)
$w_c + w_w$ (μm)	100	100	100	100	90
H (μm)	280	320	287	302	400
t_{Si} (μm)	489	533	430	458	519
P (psi)	15	15	17	31	53
f (cm^3/sec)	1.86	4.7	6.5	8.6	14.2
$\dot{q}$" (W/cm^2)	34.6	181	277	790	1309
R (cm^2-°C/W)	.280	.110	.113	.090	.083

Table 3-5: Maximum (downstream) thermal resistance of various silicon microchannel heat sinks at maximum tested power.

water decreases with increasing temperature. Thus somewhat more coolant flows to the hotter areas than is the case with the plate-fin heat sinks. This is the only significant advantage of the pin-fin structures. At low power levels the thermal resistance is linear and the performance of pin-fin and plate-fin heat sinks is virtually identical.

One sample was powered up to the subcooled boiling region. A very slow (f = 0.35 cm^3/sec), constant flow rate was used. Fig. 3-18 is a plot of the surface temperature vs. power. Beyond point "A" the output exhibited a steady "chugging" sound as pulses of steam were ejected. As the power was increased and the sample brought deeper into the boiling regime, the surface temperature began to fluctuate substantially. For a portion of the curve, the differential thermal resistance $\partial T_{surface}/\partial Q$ is actually **lower** than at lower temperatures, presumably because there is a local enhancement of heat transfer at the prime surface due to agitation [20]. However, this effect is eventually overridden as we approach the burnout power density, which is estimated by extrapolating the temperature to a vertical asymptote. This is perhaps partly due to the covering of the channel walls by vapor bubbles, and partly to the diminishing thermal conductivity of the silicon with temperature. (The burnout power level in Fig. 3-18 would be much higher if the flow rate were increased to the design value of $\sim$10

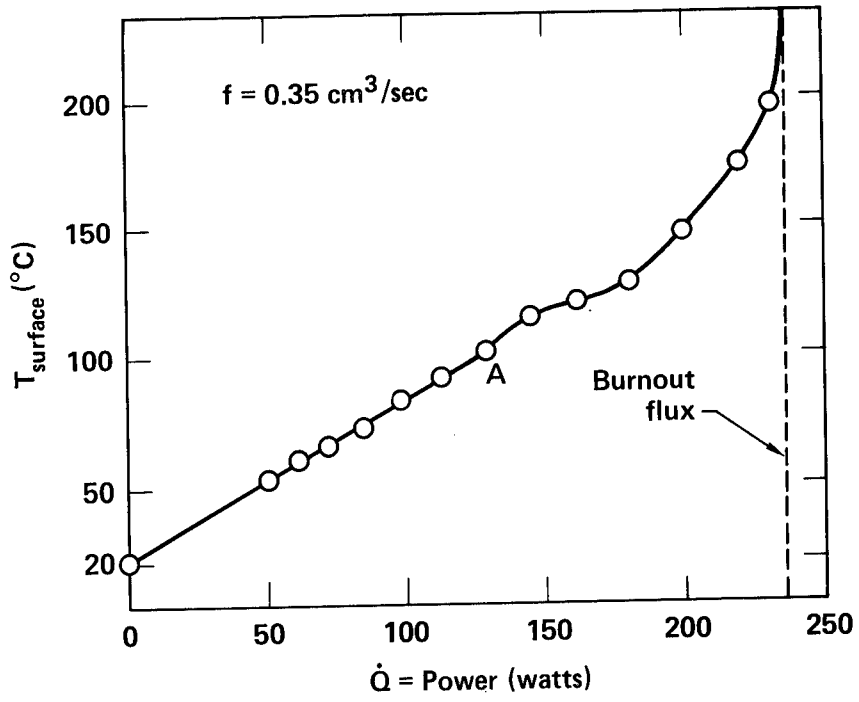


Figure 3-18: Temperature vs. power of a heat sink through the boiling regime.

cm^3/sec.) Energy calculations suggest that 15% of the water is being vaporized at burnout in this example, which is surprisingly high. Thus microscopic laminar-flow heat sinks can be safely operated well into the subcooled boiling region, just like macroscopic turbulent-flow heat sinks [20], although as a practical matter few integrated circuits could tolerate the sustained temperatures above 100°C which we have demonstrated. These results may be more useful when dealing with cryogenic coolants such as liquid nitrogen, which would normally require operation in the subcooled boiling regime.

The procedures for **predicting** the thermal performance of a heat sink have been covered in detail in Chapter 2, but we mention them here as they apply to our experiments. The normalized substrate (resistor) thermal resistance R is predicted to be

$$R(x) = LW[T(x) - T_A]/\dot{Q} = R_{cond} + R_{conv}(x) + R_{cal}(x) \quad , \tag{3.2}$$

where

$$R_{cond} = (t_{SiO_2}/k_{SiO_2}) + (t_{Si} - H)/k_{Si} \, ,$$

$$R_{conv}(x) = D/k_{H_2O} Nu_x \alpha\eta \, , \text{ where } \alpha\eta = (p_{prime} + \eta_f p_{fin})/(w_c + w_w) \, ,$$

$$R_{cal}(x) = xW/\rho Cf \text{ (optimistic theory) or } R_{cal} = xW/\eta\rho Cf \text{ (conservative theory).}$$

f is evaluated as described above. x is referenced to the upstream end of the heater resistor.

K_{H_2O}, K_{Si}, and $(\rho C)_{H_2O}$ are evaluated at the appropriate average local temperature. The silicon channel perimeters p_{prime} and p_{fin} are evaluated for the cases of sawn or etched grooves, as appropriate (Fig. 3-19). η_f is evaluated from Eq. (2.13). Nu_x is evaluated from Fig. 2-4 for an equivalent rectangular aspect ratio. A correction factor is then applied for thermal entrance effects, using Fig. 2-3. In the case of large power fluxes (and hence high temperatures), Nu_x is also corrected for the nonlinear viscosity effects, as described in Section 2.2.4.

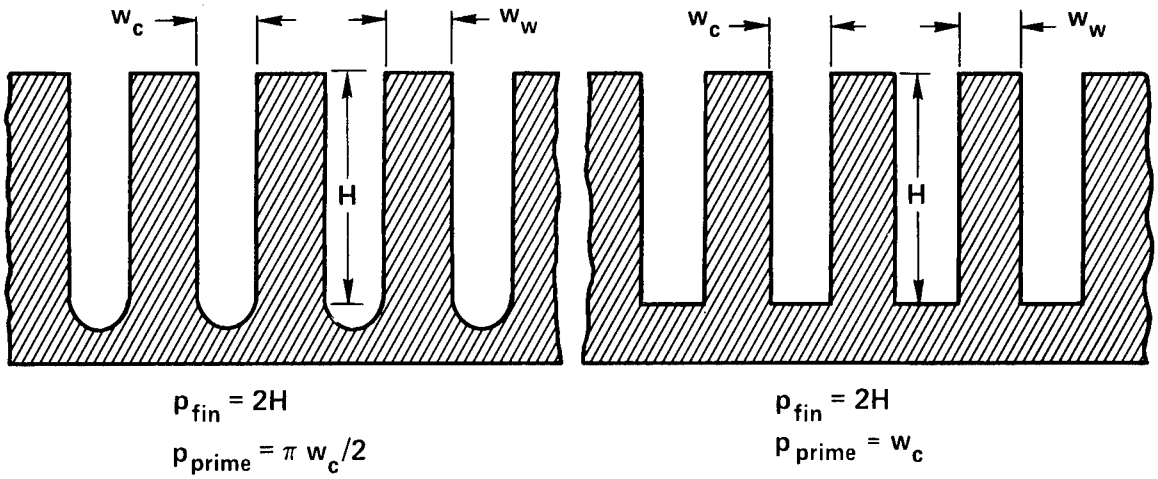


Figure 3-19: Model of sawn grooves (circular bottoms), etched grooves (square bottoms).

A series of experiments was performed in which R was mapped as a function of downstream position x, for a fixed power $\dot{Q}$ and flow rate f. This provides a sensitive test of the correctness of our analyses. Fig. 3-20 plots theory and experiment for sample 82A26A2 at several different flow rates and positions, correlated in terms of the dimensionless axial position $x^* = x/(D \cdot Re \cdot Pr)$, which is proportional to x/f. The agreement with theory is quite good, in view of the fact that no adjustable parameters were used. The linear slope for large x^* is due simply to caloric heating, and matches exactly that expected. The nonlinear behavior at small values of x^* is due to the developing thermal boundary layer and has the expected shape. Although we have not probed deeply into the low-x^* region, there is no reason to expect significant deviations from the theoretical curve in Fig. 3-20. The conservative and optimistic theories are virtually identical for this sample and so only a single theoretical curve is shown.

Fig. 3-21 plots thermal resistance for another sample (82A26C1), as a function of downstream position x. Here the optimistic and conservative theories are both shown and, as

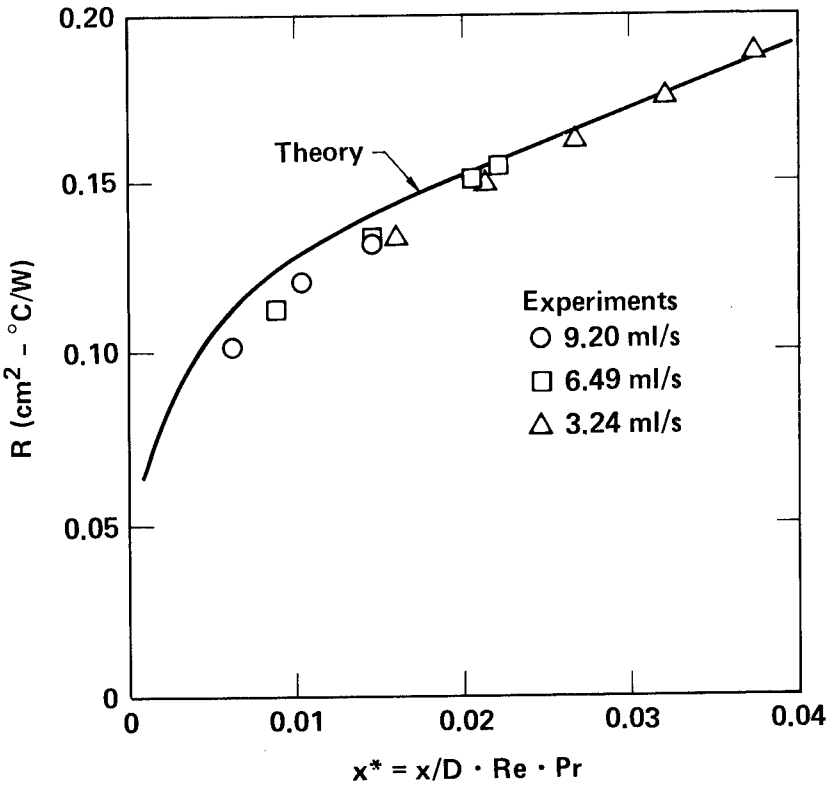


Figure 3-20: Thermal resistance R vs. dimensionless position x^* for sample 82A26A2.

expected, the data generally fit between them. The measurement errors are approximately equal to the size of the plotted points. The drooping near x = 10 mm (the downstream edge) is due to thermal spreading at the edge, as calculated in Section 2.2.5. In this sample, the thermal resistance is primarily caloric (a lower-than-optimal flow rate).

To summarize, microscopic laminar-flow heat sinks have been fabricated having demonstrated performance in good agreement with the theory of Chapter 2. Peak thermal resistances of as low as 0.083°C/W for a $(1\ cm)^2$ heater have been measured. Although the channel length L_s was always at least 1.4 cm in our samples, this could be scaled down to yield performance benefits proportional to $L_s^{-1/2}$. The performance of pin-fin and plate-fin structures were very similar for a given flow rate, except that the pin-fin structures allow more water to flow near hot spots due to viscosity reduction. The heat sinks can be operated in the subcooled boiling region without problems and exhibited a large burnout margin (a factor of 2 beyond the initial boiling).

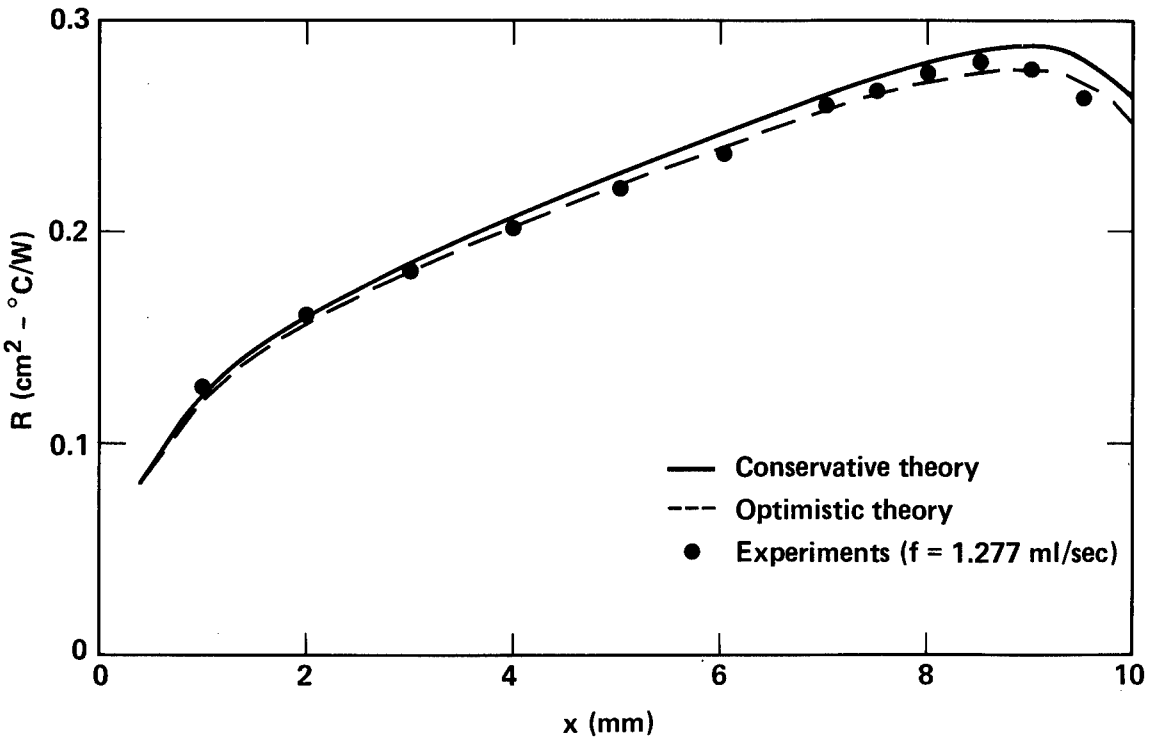


Figure 3-21: Normalized thermal resistance as a function of position for sample 82A26C1.

3.2.5. Long-Term Reliability

We believe that the use of filtered, deionized water will allow very long (multi-year) heat sink lifetimes, free from significant fouling or erosion. Heat-exchanger fouling (an increase in flow friction) is traditionally caused either by the clogging of channels with particles or by oxidation of the channel walls (i.e., formation of a "scale"). Particles could be eliminated simply by filtering the water; for example, a 15-μm filter should be adequate for protecting 50-μm channels from fouling. Oxidation is not a problem with silicon, because the oxidation rate at room temperature is unmeasurably small once an initial passivating layer $\simeq$1 nm thick has formed.

Channel erosion (a decrease in flow friction) is a more likely problem and there are two mechanisms by which it might occur: chemical (etching) or physical (ablation). Chemical attack would only be a problem if the water was significantly alkaline (e.g., containing KOH) or contained large amounts of fluoride ions. Neither situation exists when deionized water is used; a silicon wafer can sit in DI water indefinitely without any measurable change in thickness. Physical erosion would occur only if the material limits (fracture stress) of the

silicon were exceeded. It is known [68] that single-phase flow will not cause erosion because the highest pressure seen by the material is the stagnation pressure $P_\infty = P + \rho v^2/2$, which in our designs is at most 50 psi (345 kPa). This is far below any material limits. Physical erosion can occur in multiphase flow, the two common causes being cavitation (rapid collapse of vapor bubbles) or physical ablation by solid particles (e.g., sandblasting). Our designs are well below the cavitation limit because $\rho v^2/2 \ll 1$ atm. Physical ablation by entrained particles remains the most likely long-term reliability problem. The details of the erosion mechanism are complex [68, 69], but substantial pipe erosion is known to occur at velocities of $v \simeq 10^4$ cm/sec. Since our velocities are typically 20 times smaller, it is not clear whether such erosion could occur, but certainly using well-filtered water should minimize the problem.

An extended experiment was performed on a heat-sink sample to determine whether any measurable erosion occurred in our designs. Deionized water (18 MΩ, 18°C), filtered to better than 15 μm, flowed continuously through sample 83MR31E (see Table 3-3 for dimensions). The supply pressure was 30 psi (207 kPa) and the flow rate was 7 cm^3/sec, corresponding to a mean channel flow velocity of 245 cm/sec. After 1000 hours of operation, no change in flow friction could be observed within the measurement error of $\pm$1.5%. Thus erosion does not appear to be a problem in our heat sinks.

Chapter 4
Microcapillary Thermal Interface: Design

4.1. Background and Prior Art

4.1.1. Solid Thermal Interfaces

In the preceding chapters, we described a new technique for removing large heat fluxes from planar integrated circuits by manufacturing microscopic heat-transfer structures directly within the IC substrate. Such a procedure raises questions of fabrication yield, reliability, and practicality. Fabrication yield may suffer because the processes involved in manufacturing the heat sinks may damage the IC. Reliability may be influenced by diffusion of contaminants such as gold from the water into the silicon. A practical consideration in multi-chip systems such as computers is the complexity of supplying each individual IC with coolant. To avoid these problems, it may be desirable to fabricate the microchannel heat sinks as a separate "cold plate" (ideally a part of the circuit card or module). The cold plate might be fabricated out of copper (rather than silicon) for maximum performance, or out of ceramic for compatibility with multilevel wiring. The unmodified chips would then be bonded to this cold plate/circuit card. The chips could be electrically connected to the cold plate/circuit card by conventional techniques such as wire bonding or, better yet, area TAB (area tape automated bonding) [70]. However, the quality of the thermal interface between the IC and the cooled board is now critical. A normalized thermal resistance of much less than 0.1 $cm^2 \cdot {}^{\circ}C/W$ is desirable in order not to significantly degrade the thermal resistance of the package.

There exist a number of conventional attachment techniques involving solid interfacial layers (eutectic die bonding, solder die bonding, or epoxy die bonding) [71] which may be adequate under some circumstances. However, each of them has limitations, especially when applied to large die sizes, high power densities, and/or multi-chip ("hybrid") packages which are being adopted in high-performance digital systems. The limitations are related, respectively, to mechanical stress, thermal resistance and voiding, and the inability to detach and remake the interface.

Mechanical stress is characteristic of all the conventional bonding techniques when a mismatch in thermal expansion coefficient exists between the silicon, adhesive layer, and the cold plate/board. This could be minimized by making the cold plate of silicon. If this is not practical, then the differing thermal expansion coefficients α will generate shear stresses when the temperature changes. If the bond was made at elevated temperatures (as is usually the case), continual shear stresses will exist at room temperature. These stresses increase with die size, and are largest at the edge of the die. For eutectic die bonds or hard solders, this edge shear stress τ can be as large as $\sim 10\times$ the tensile stress $\sigma = \varepsilon_{th}E$, where ε_{th} is the thermal strain. This can cause fracture of the die and delamination when die size exceeds about 1 cm, particularly when a metallic substrate is used [72]. To circumvent this problem, "soft solders" or thermally conductive epoxies are often used for larger die sizes; the soft solder has a low shear modulus G and hence acts as a strain buffer. The maximum edge shear stress in a large die has been found to be [73]

$$\tau_{max} \simeq \varepsilon_{th} \cdot [G/(t_{solder}/E_1 t_1 + t_{solder}/E_2 t_2)]^{1/2}$$

where t_{solder} is the solder or epoxy thickness and t_1, t_2, E_1, and E_2 are the thickness and elastic constants of the chip and cold plate. Evidently one desires that the solder layers be thick in order to reduce the shear stress on the chip, but this is contrary to our desire to minimize thermal resistance. For a typical soft solder, the thermal conductivity is $k \simeq 0.5$ W/cm-K; for a good thermally conductive epoxy, $k \simeq .04$ W/cm-K. More seriously, the use of a soft solder does not solve the mechanical stress problem for large dice, but merely trades it for the problem of low-cycle solder fatigue [74]. This is caused by repeated thermal cycling, which causes plastic deformation and eventual work-hardening. The bond will ultimately develop a fatigue fracture, and Fig. 4-1 shows data from Ref. [75] predicting failure in fewer than 400 cycles for very large chips. Even before the joint fails mechanically, internal voids and cracks are created by the thermal cycling, which significantly increase the joint thermal resistance [72]. In fact, an increase in thermal resistance is the most sensitive predictor of fatigue failures in die attachments [76].

In addition to the mechanical stress problem, conventional die attachments have thermal limitations, particularly for large chips. The problem of "voiding", in which the bond contains voids due to trapped gas, outgassing during the reflow, incomplete wetting, entrapped flux, metallurgical incompatibilities [77], nonplanarity of the die, or low-cycle fatigue, is well known and remains a problem in present-day IC packaging. Such voids increase the thermal resistance at local areas, causing hot spots. The problem is expected to become more critical as die sizes and power densities are increased [78]. Both eutectic die bonds and solder die

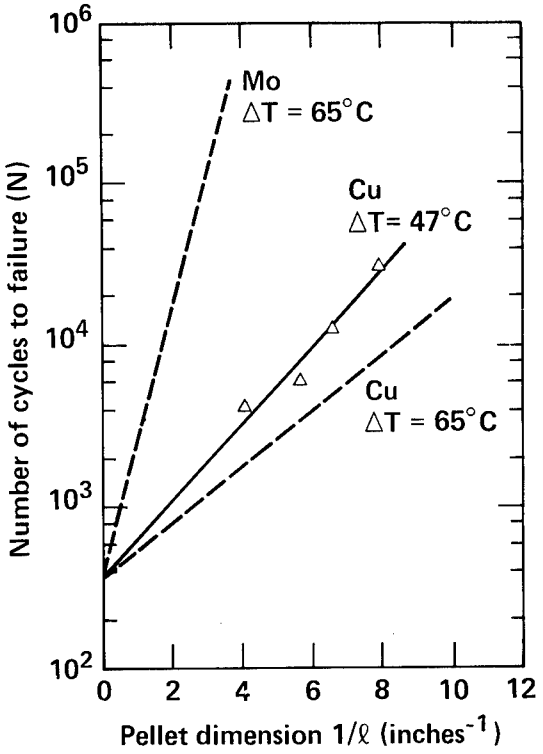


Figure 4-1: Thermal fatigue failure curves for silicon mounted on molybdenum or copper (from Lang et al [75]).

bonds are prone to voiding, and epoxy bonds are worst of all since outgassing occurs during the curing process.

The conventional attachment techniques do not generally allow the detachment of a "bad" die and the subsequent attachment of a "good" die to substrates. Irreversible metallurgical reactions occur in the thin films or substrates used in die bonding, so the surface cannot be properly repaired for a reattachment, even if the die could be successfully detached. While this is not a problem in single-chip packages which are discarded if the chip is bad, it would be a severe problem in multi-chip modules, where the failure of a chip would make an entire module useless.

Finally, the inability to detach and remake die bonds also implies that they cannot be used for testing of dice **prior to packaging.** A reversible high-thermal-conductance thermal interface allowing full-power testing of large, high-power-density ICs has not been heretofore developed, to our knowledge.

To summarize, conventional solid thermal interfaces have limitations due to mechanical

stresses, high thermal resistance (especially from voiding), and nonreusability. These limitations may make them unsuitable for use in large-area, high-power, multi-chip modules which may constitute future high-performance computers. We would therefore like to develop a thermal interface technology for high-power, large-area chips, which provides low thermal resistance ($R \ll 0.1$ cm^2-°C/W) over the entire area, low (preferably zero) mechanical stress, is detachable, and which allows the die to be nondestructively tested at full power prior to final packaging on a cooled multi-chip module. Note that, unlike for power semiconductor devices, good **electrical** contact is **not** a requirement. In virtually all VLSI technologies, the substrate current is negligible (due only to leakage). Often such substrate contacts are achieved by a separate bonding pad to the front (circuit) side of the chip, but even with a direct substrate contact its electrical resistance need not be very low.

4.1.2. Gaseous Thermal Interfaces

One way to achieve a low-stress, reusable thermal interface would be to use a gas as the thermal interface medium. A high thermal-conductivity gas such as helium (k_{He} = .00153 W/cm-K) would be preferable. However, in order to achieve sufficiently low thermal resistance a very small gap is required, even using helium gas. For example, an "average" gap of less than 0.76 μm is required to achieve $R < 0.05$ cm^2-°C/W. The term "average" is used loosely here, for in most cases microscopic contact would exist between the surfaces at a large number of points. There is copious literature on the thermal resistance of such contacts [79, 80, 81], including theoretical lower bounds [82]. Generally the minimum possible thermal resistance is the order of conduction across a phonon mean free path length in each of the materials involved. The molecular mean free path in the gas can also be a limiting factor but this can be remedied simply by increasing the pressure. These theoretical lower bounds are well under 0.05 cm^2-°C/W, so we need not be concerned with them in this study. All we need know is that if the maximum gap between the surfaces is less than 0.76 μm, then a thermal resistance of less than 0.05 cm^2-°C/W through helium gas is assured. This could be achieved if both surfaces were optically flat and dust-free.

In practice integrated circuits are warped due to stresses in their various thin films, especially those deposited or grown at high temperatures (e.g., thermally grown SiO_2). Substantial mechanical pressure would be required to achieve an effective gap of less than 0.76 μm over the entire contact area. Furthermore, for large chips, this pressure would have to be fairly uniformly applied over the die area to ensure that the warpage is flattened out at all points and to avoid excessive localized stresses. Such mechanical loading is an accepted

technique for achieving good thermal contact between surfaces, but may not be practical in computer applications due to the need for electrical access to the many bonding pads on the front surfaces of the chips and the relative fragility of the deposited films. Moreover, the need for a virtually perfect hermetic seal to confine the highly mobile helium atoms is an adverse practical consideration.

IBM Corporation has pioneered the use of helium gas as a thermal interface medium in their "thermal conduction module" (TCM) [13, 14]. In this structure, a piston rests on the back of each chip to achieve thermal contact. Electrical contact is made by soldering the chips face down onto an array of "C4" solder balls. This arrangement limits the piston force which may be applied, and also is prone to chip tilt. To circumvent these problems, IBM uses a convex spherical piston head to ensure contact near the center of the chip, at the expense of relatively poor contact elsewhere. The interfacial thermal resistance is 2.9°C/W over a chip area of .209 cm^2, hence R≃0.6 cm^2-°C/W. An additional thermal interface resistance of 2.15°C/W exists between the piston and cylinder walls. If the problems of chip tilt and warpage could have been eliminated and smooth flat surfaces used (RMS roughness = 0.4 μm), a 6-fold improvement would have been achieved [13].

To summarize, helium gas thermal interfaces between very flat surfaces are reusable, free from shear stress, and have the potential for excellent thermal performance. However, wafer warpage and/or tilt and mechanical loading restrictions presently prevent it from reaching its theoretical potential. Furthermore, the package hermeticity requirements are stringent, owing to helium's high diffusivity through most materials.

4.2. Principles of Liquid Thermal-Conduction Interfaces

4.2.1. The Basic Idea

This work explores the use of a **liquid** as the thermal interface medium, rather than a gas or a solid. As with a gaseous thermal interface, the contact would be free from shear stress and detachable and reusable. The thermal resistance of such an interface would be determined by the liquid's thermal conductivity and by the interfacial gap. With the exception of liquid metals, the thermal conductivity of most liquids is 1 or 2 mW/cm-K, i.e., comparable with that of helium gas. Thus the interfacial gap should be limited to 1 or 2 microns to achieve normalized thermal resistances of less than 0.1 cm^2-°C/W. The challenge, then, is to design a liquid thermal interface which:

1. Guarantees that the liquid uniformly wets the contact surfaces without any voiding.

2. Corrects for wafer warpage, to achieve a minimum-thickness gap (Preferably this is achieved without external mechanical loading.)

3. Is reliable (i.e., chemically compatible and long-lived).

These problems are nontrivial, and so far liquid thermal interfaces have not been used in IC packaging at the chip level. The closest thing is the use of various "thermal grease" compounds, consisting of silicone pastes filled with small thermally conductive particles, which are used to interface IC packages to heat sinks. These can yield heat transfer coefficients of up to 10 W/cm^2-K when the surfaces are clamped under heavy compressive loads [83]. This may not be practical at the chip level, where many electrical contacts to the front of the wafer are required.

Our approach is to make use of classical surface tension and capillary theory [84] to engineer a reliable, void-free, high-conductance thermal interface. Fig. 4-2 illustrates our technique for doing this.

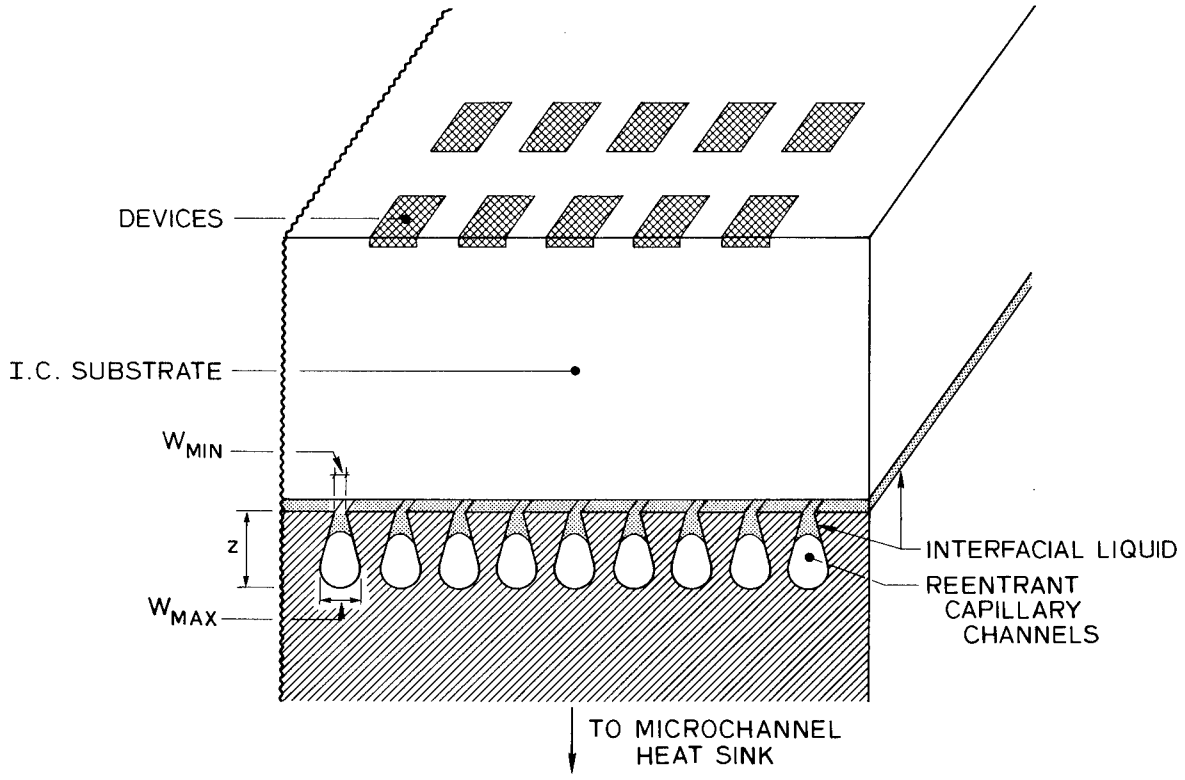


Figure 4-2: Microcapillary thermal interface concept.

A limited amount of interfacial liquid partially fills a set of long, parallel, open-ended

microscopic reentrant capillaries which cover the entire surface of the substrate. (A reentrant groove is one which is narrower near its top than near its bottom). In this example, the microcapillaries are fabricated in the heat sink substrate; equivalent results could be achieved by fabricating them in the chip instead. There are several reasons why the liquid supply is limited so as to only partially fill the capillaries.

First, the capillaries can then act as reservoirs for excess liquid while maintaining a minimum-thickness interface. If the volume of the interfacial gap region changes due to a slight thermal flexure of the chip, or the volume of the liquid changes due to thermal expansion, interfacial liquid can rapidly move in or out of the grooves to compensate for this volume change. This equilibration occurs very rapidly (much faster than the thermal time constants of the system); hence there is no danger of an area being deprived of liquid as a result of thermal cycling.

Second, the capillaries enable trapped air to escape out the open ends, so large voids cannot occur. The only voids which could exist would be ones which were small enough to fit between adjacent capillary grooves. Such voids are much smaller than the chip thickness and hence would not measurably affect the thermal resistance seen by the chip. Moreover, such voids could only be stable at a local maximum in the gap (see Fig. 4-8, page 99), whereas the fabrication process for the reentrant grooves tends to preclude the existence of such local maxima.

Third, the geometry enforces a well-defined attractive (suction) force between the two surfaces due to the liquid's surface tension, which can significantly reduce wafer warpage (hence reduce the interfacial gap thickness) without any external mechanical loading. Assuming near-zero contact angle, the suction pressure (equal to the negative hydrostatic pressure in the liquid) is $P_o \simeq \gamma/r$ where γ is the surface tension of the liquid and r is the radius of the liquid meniscus. For high-aspect-ratio grooves, $r \simeq w_m/2$, where w_m is the width of the groove at the level of the meniscus. Using 2-μm-wide grooves, this pressure would be 0.37 atm in silicone oil ($\gamma = 0.37$ ergs/cm^2), 0.74 atm for water ($\gamma = 74$), and $\sim$5 atm for liquid metals such as Hg or Ga-In ($\gamma \simeq 500$). Obviously, additional mechanical loading could be applied so as to further reduce the interfacial gap beyond that achieved by capillary suction alone.

It might seem that the suction pressure could be increased without limit simply by scaling down the capillary dimensions, but this is only the case for perfectly flat (zero gap) surfaces.

If a finite gap w_{edge} exists at the edge of the chip, that edge will de-wet as soon as the suction pressure P_o exceeds $2\gamma/w_{edge}$, and the liquid will recede until it reaches a region where the gap is less than $2\gamma/P_o$. Thus it is desirable to design the groove widths to be slightly larger than the maximum expected gap between the surfaces at the relevant contact pressure. If this is not done, we are no longer guaranteed a void-free interface.

One intriguing property of this configuration is that the capillary suction force will exist without regard to the outside ambient pressure. For example, the suction force would still exist in a high-vacuum environment, provided a low-vapor pressure liquid such as silicone oil were used. The liquid is thus under a negative absolute pressure (hydrostatic tension), but this is no problem because clean (particulate-free) liquids have tensile strengths of hundreds of atmospheres [85, 86]. Moreover it can be readily verified that the liquid cannot rupture inside the capillaries, because this would require that the "critical bubble radius" [87] for vapor nucleation exceed the capillary width. This could be a very effective technique for flattening and heat-sinking wafers in high-vacuum environments such as evaporators, ion implanters, or electron-beam equipment, without permanently bonding the wafers (traditionally a very difficult problem).

Prior to assembly, the liquid must be applied to the planar surface (i.e., the chip or wafer), either by spinning or as a single droplet. In the latter case several seconds are required for the liquid to distribute itself, during which time all voids will be expelled. The chip may be later detached either by pulling (applying a tensile force in excess of the suction force) or by flushing the area with excess interfacial liquid, thus filling the groves and causing the capillary force to vanish.

To ensure the ability of liquid to fully equilibrate between adjacent grooves, it may be desirable to fabricate tunnels between such grooves (Fig. 4-3a). An even better configuration would be a two-dimensional grid of reentrant grooves, as shown in Fig. 4-3b. This would allow free flow of liquid in both directions. However it is more difficult to fabricate this structure and we did not investigate it experimentally.

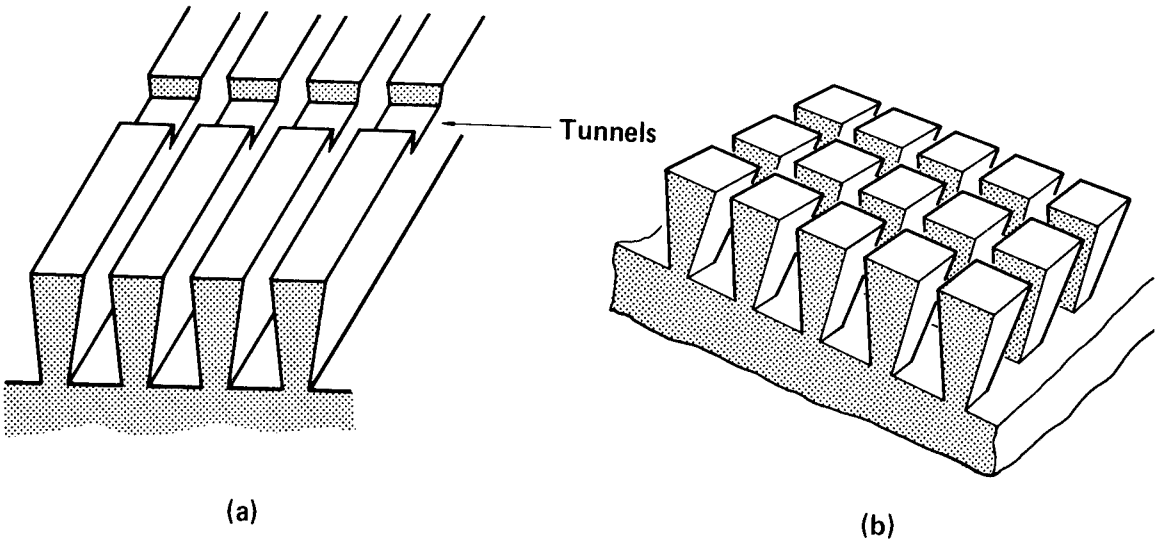


Figure 4-3: a) Tunnels between adjacent capillaries facilitate global equilibration of liquid. b) Proposed two-dimensional array of reentrant grooves.

4.2.2. Reentrant Grooves

The reentrant shape of the capillaries is crucial to the initial distribution and the subsequent void-free stability of the liquid layer. The array of reentrant capillaries is hydrostatically stable in only one configuration, namely that shown in Fig. 4-2, in which the liquid is uniformly distributed among the grooves. To see this, suppose the liquid level in a capillary is perturbed; then the radius of curvature of the meniscus will change, resulting in a change in the local hydrostatic pressure with respect to its neighbors. This pressure change will force some liquid to flow to or from neighboring channels so as to restore the original liquid level (Fig. 4-4a). In contrast, if the grooves had conventional taper, the perturbed grooves would end up either completely filled or completely empty, depending on the initial perturbation. Alternatively, most of the liquid could migrate to the bottoms of the grooves (where it is narrower and hence the total surface energy is minimized), becoming discontinuous from the interfacial layer. The result of using normally tapered grooves would thus be a highly nonuniform, discontinuous distribution of liquid, where only a small, highly disconnected fraction of the liquid was in the interface. No reservoir effect exists to allow the chip to rapidly correct for slight changes in the gap volume or liquid volume. Furthermore the liquid is much more prone to evaporation because "hot" surfaces (those in contact with the chip) are exposed, whereas in the reentrant groove configuration of Fig. 4-2 the only place where menisci are exposed is deep within the heat sink substrate where the temperature is near ambient.

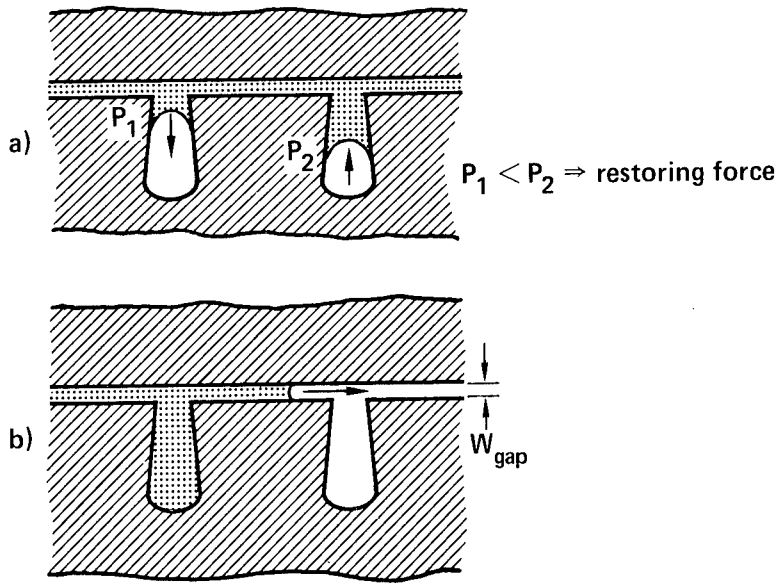


Figure 4-4: Capillary stability properties of long reentrant grooves.

The open-ended reentrant capillaries will restore a uniform distribution of liquid even in the event of a large initial perturbation (such as applying the liquid as a droplet). For example, suppose a completely filled reentrant capillary is adjacent to a completely empty one. Provided the end of the filled capillary is open, the capillary force of the gap will initially draw liquid from the end of the filled capillary, followed by a dewetting of the bottom of the groove, until both grooves are equally partially filled (Fig. 4-4b). It is not required that the the reentrant grooves be smoothly tapered as sketched in Fig. 4-2. An abrupt necking as shown in Fig. 4-5 would also provide adequate stability.

An experiment was performed to confirm that reentrant microcapillaries do indeed generate a uniform liquid distribution. A parallel array of grooves having a depth slightly less than the wafer thickness and a width of 30 μm were fabricated in a silicon wafer by precision sawing. The sawing process produces a conventional (as opposed to reentrant) taper in the grooves. The grooved face was then anodically bonded to a Pyrex plate, and the silicon was etched in 1:3:4 HF:HNO_3:HAc until the bottoms of the grooves were exposed. Now the exposed grooves are reentrant when viewed from that side. A measured droplet of silicone oil was applied to an optically flat glass slide, which was then mated with the grooved wafer. The mating surfaces exhibited a capillary attraction, as expected. By sighting directly down the grooves in an optical microscope with transmission illumination, we directly observed the menisci (Fig. The light areas are where there is no oil; the dark areas immediately above

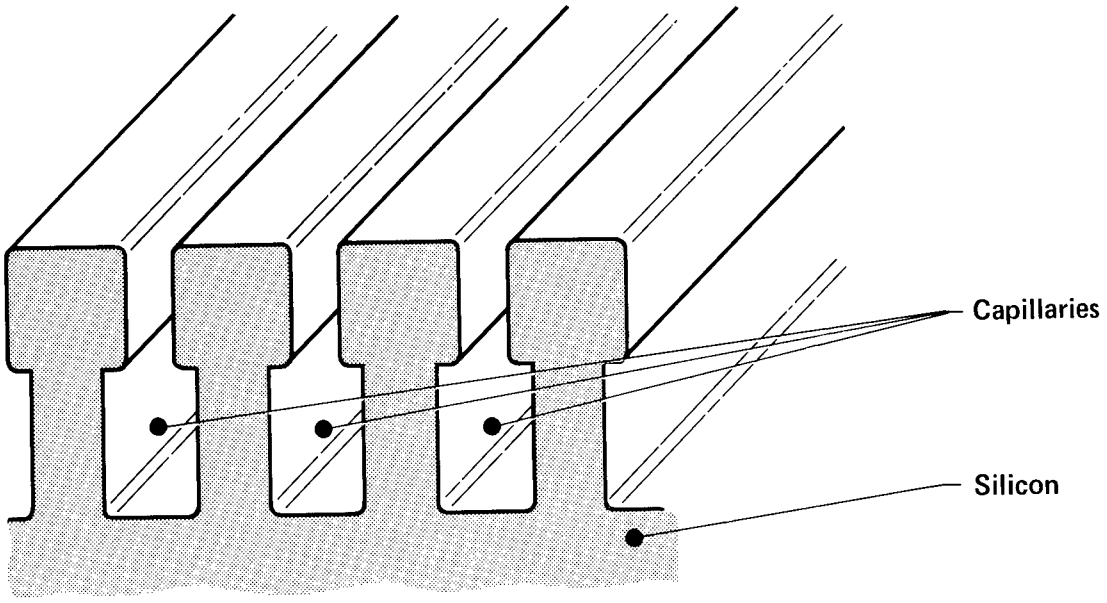


Figure 4-5: Abruptly-tapered reentrant capillary grooves would also be acceptable.

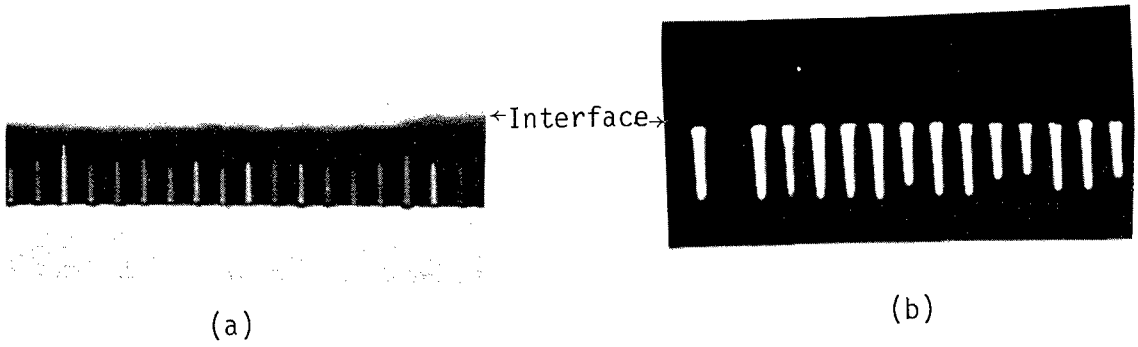


Figure 4-6: (a) Verification of reentrant capillary stability using 30-μm wide, 400-μm deep reentrant grooves. The menisci show that the interfacial oil (dark portions of grooves) congregates near the interface. (b) With normally-tapered grooves, the oil congregates away from the interface (at the bottoms of the capillaries).

indicate the presence of a column of oil. Note that the menisci are clearly semicircular and all at approximately the same level. (The slight differences are due to slight variations in capillary width due to manufacturing tolerances.) The liquid has congregated near the interfacial layer, as expected.

In contrast, Fig. 4-6b shows the results of a similar experiment using normally tapered

grooves. Here nearly all the liquid has congregated at the narrow bottoms of the grooves, away from the interfacial layer. Only small isolated pockets of liquid remain near the interface.

Although these demonstrations were performed on grooves 10 times larger than were ultimately used, the principles are scale-invariant. An additional experiment, which demonstrates the dramatic differences in capillary action between reentrant taper and normal taper, was performed on grooves identical to those used in the experiments of Chapter 5. Here, a drop of photoresist was applied and then spun at 4000 RPM for 30 sec on a wafer containing reentrant grooves (2 to 4 μm wide). The result is shown in Fig. 4-7a. The photoresist congregated at the tops (narrow portions) of the capillaries before the solvent evaporated, nearly planarizing the surface. In contrast, Fig. 4-7b shows the results of spinning photoresist on normally tapered grooves. Here the photoresist congregated at the bottoms (narrow portions) of the grooves before drying, leaving the surface almost untouched and not at all planarized. This demonstrates how capillary forces at micron scales can have a profound effect on liquid behavior, even when spinning at 4000 RPM!

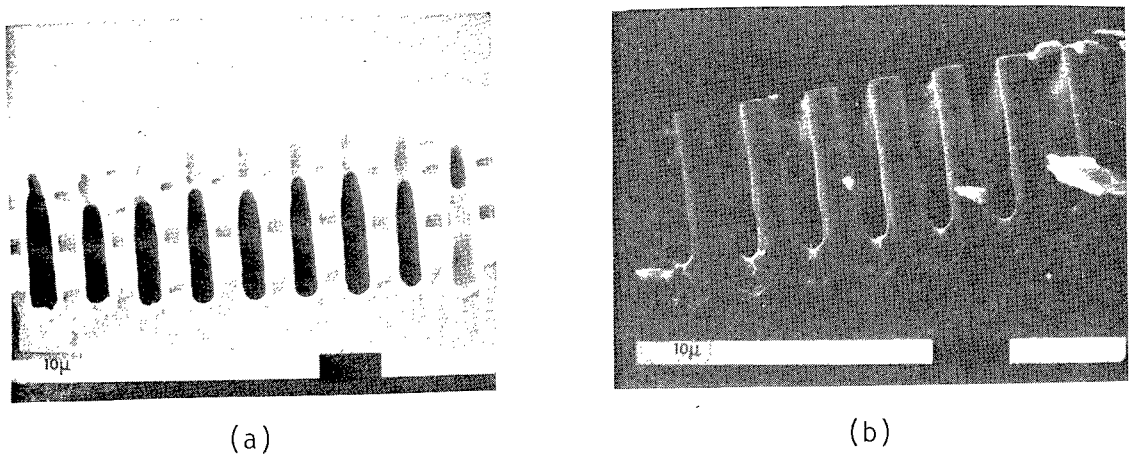


Figure 4-7: SEM of photoresist spun on: a) reentrant microcapillaries, and b) conventional-taper microcapillaries.

It should be noted that it would be very difficult to achieve a minimum-thickness void-free liquid thermal interface without using microcapillaries. The obvious approach would be to coat two planar surfaces with liquid and bring them into contact. If this were done, the liquid interface would usually have too much interfacial liquid at first (too thick a conducting layer). If the surfaces were squeezed together to expel the excess liquid, then an extremely small volume of liquid would remain in the interfacial region. A slight perturbation in volume due to

thermal flexure, thermal contraction, or change in mechanical loading could leave an insufficient amount of liquid to fill the gap. Then the remaining liquid would redistribute so as to produce voids at any local maxima in the gap, which could be at the center of the wafer. Such voids would ruin the thermal performance of the interface (Fig. 4-8).

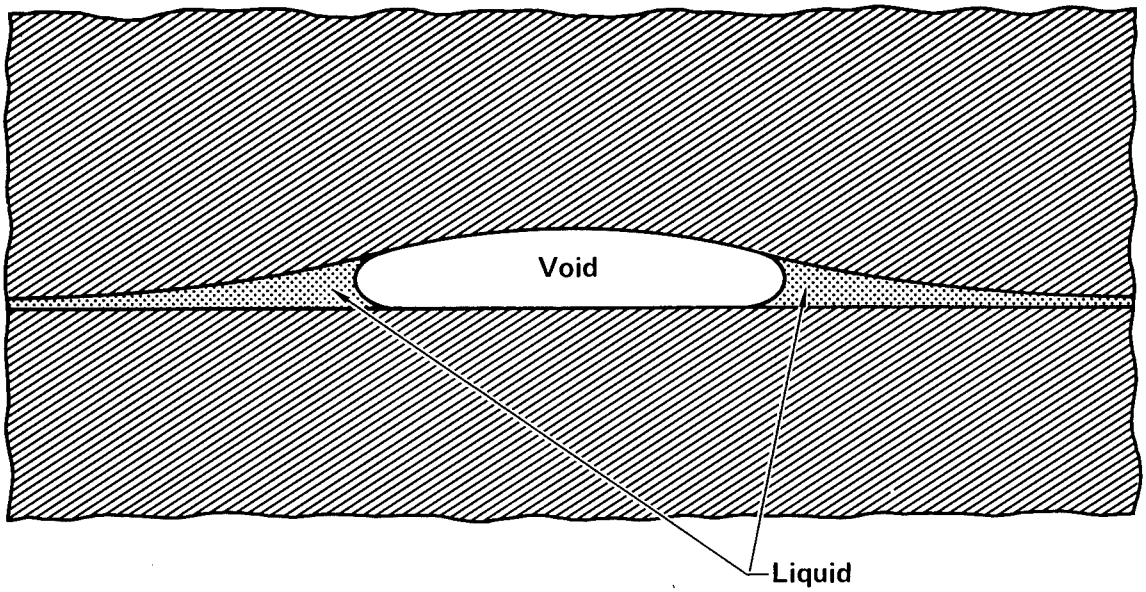


Figure 4-8: Voids would congregate at local maxima in the gap between planar surfaces.

4.3. Interfacial Gap

The heat-sinking performance (thermal conductance) of the microcapillary thermal interface will be determined by the average gap between the surfaces and by the thermal conductivity of the interfacial liquid. In an ideal situation, the mating surfaces would be perfectly smooth, optically flat, and dust-free, implying zero gap and hence essentially zero thermal resistance. In real interfaces, the surfaces are smooth but not necessarily flat, primarily because of process-induced wafer warpage. Furthermore, there will assuredly be some entrapped particles in the interface; the amount distribution of sizes depends primarily upon the cleanliness of the fabrication environment. Counteracting these effects, the suction pressure from the microcapillaries will draw the surfaces together. We therefore wish to analyze the interfacial gap as a function of warpage, entrapped dust, and mechanical pressure. Our analysis will be "worst-case"; microscopic roughness of the surfaces and plastic deformation of the mating surfaces will result in lower thermal resistance than our simple gap calculations would indicate.

4.3.1. Wafer Warpage

We first consider how much warpage is expected in the silicon chip or wafer. Prior to processing, the warpage is extremely small; typically the radius of curvature ρ of a new, polished silicon wafer exceeds 50 meters. This can be related to the maximum bow w_{bow} for a circular die of radius R: $w_{bow} \simeq R^2/2\rho$, which is less than 1 μm for a 2-cm diameter die. As will be shown shortly, this is well within the maximum tolerable bow.

The real problem is that wafers become warped as a result of stresses induced in the silicon wafer and its deposited thin films during IC fabrication. For example, the thermal coefficient of expansion of silicon is substantially larger than that of amorphous SiO_2. Thus a high-temperature thermal oxidation of silicon would produce a significant tensile stress in the silicon at room temperature, elastically warping the wafer. Jaccodine and Schlegel [88] have studied this case and have shown that the measured strain agrees with that expected from elastic plate theory. Similar results have been found for silicon nitride films [89].

There is evidence that **plastic** deformation can also contribute to wafer warpage [90]. The differential thermal expansion mismatch from the thin films is not normally sufficient to cause plastic deformation, even at high processing temperatures [88]. However, excessively fast temperature cycling of the wafers (i.e., a "fast pull" from the furnace) can produce momentarily large radial thermal stresses sufficient to plastically deform (and hence grossly warp) the wafer [90]. However, such warpage also causes severe problems in subsequent lithographic steps due to highly nonlinear in-plane distortion. For this reason, the sources of gross plastic deformation are generally eliminated in a modern IC process. Yau [91, 92] has shown that a modern nMOS process produces wafer warpage entirely consistent with purely elastic deformation due to thin-film stresses. He further showed that ρ>10 m in all cases. We shall therefore assume that the magnitude of the local radius of curvature ρ is always greater than 10 m, as a worst-case estimate of wafer warpage. (We do not necessarily assume that ρ is a constant across the wafer; the local curvature could conceivably vary slightly from point to point).

4.3.2. Smooth Plate Deflection Theory

We now consider the effect of a uniform mechanical suction pressure P_o on the interface gap. In our analysis, we assume the back of the wafer to be perfectly flat, and the heat sink substrate to be smooth but warped; obviously this is equivalent to a warped wafer in contact with a flat substrate, since this is a linear (small-deflection) problem.

Consider the case in which an initially flat circular chip or wafer of radius R and thickness t is pressed against a concave substrate having radius of curvature ρ (Fig. 4-9).

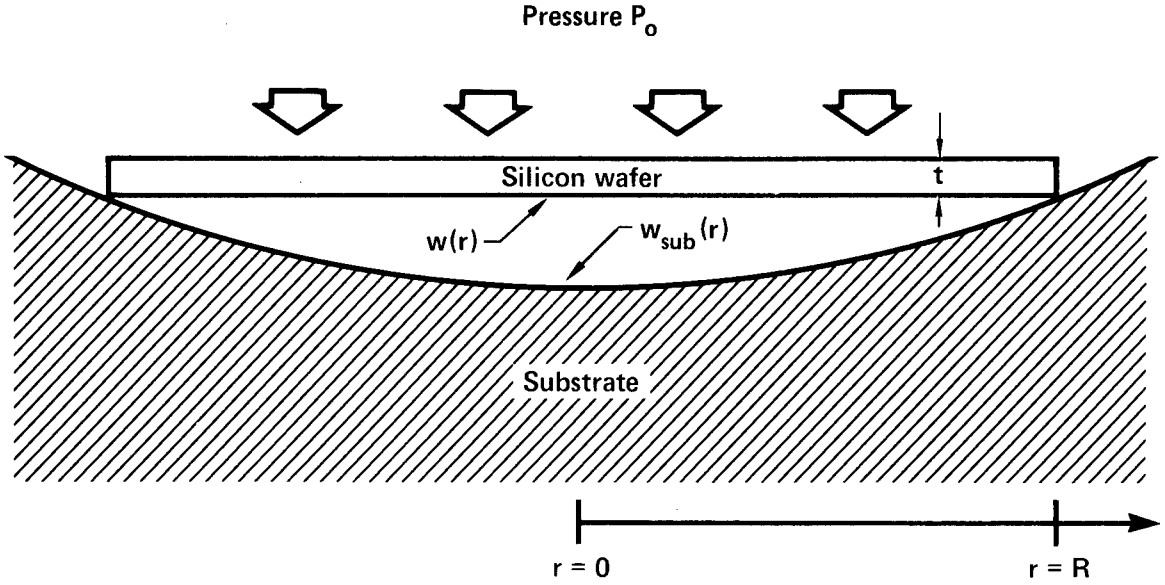


Figure 4-9: Deflection of a wafer against a concave substrate under uniform pressure P_o.

The depth w of the wafer at the circumference is defined to be zero (w(R) = 0); the concave surface of the substrate w_{sub} is therefore defined by the equation

$$w_{sub}(r) = \sqrt{\rho^2 - r^2} - \sqrt{\rho^2 - R^2} \simeq (R^2 - r^2)/2\rho \quad \text{for } R \ll \rho.$$

In view of the high aspect ratios of silicon chips or wafers, their deflection may be modeled by classical elastic plate-deflection theory [93]. Initially the wafer is "simply supported" at the edges (r = R). The elastic plate deflection equation for the wafer is:

$$\nabla^4 w = P_o/D \tag{4.1}$$

where $D = Et^3/[12(1-\nu^2)]$. Here E is the elastic modulus of the silicon ($E = 2\times10^{12}$ dynes/cm^2) and ν is Poisson's ratio for silicon ($\nu = 0.09$) [94]. The boundary conditions are:

$$w = 0 \text{ and } \partial^2 w/\partial r^2 + (\nu/r)\partial w/\partial r = 0 \quad \text{at } r = R \text{ (simple support).} \tag{4.2}$$

Solving Eqs. (4.1) and (4.2) yields the plate deflection as a function of r:

$$w(r) = \frac{P_o}{64D} \cdot \left\{ r^4 - 2\left(\frac{3+\nu}{1+\nu}\right) R^2 r^2 + \left(\frac{5+\nu}{1+\nu}\right) \cdot R^4 \right\}$$

The maximum plate deflection occurs at the center (r = 0), where for silicon,

$$w(0) = 3(1-\nu)(5+\nu)P_o R^4/16Et^3 = (P_o R^4/t^3)/(2.3\times 10^{12}\ \text{dynes/cm}^2)\,.$$

This solution is valid provided the pressure is not sufficient to bring the surfaces into contact at any point other than the edges. When such contact does occur, the boundary conditions change because additional support is now being provided.

The **gap** between the deflected wafer surface w(r) and the uniformly concave substrate $w_{sub}(r)$ is minimized at the center (r = 0), i.e., $w_{gap}(r) \equiv w_{sub}(r) - w(r)$ has its minimum at r = 0. The case in which $w_{gap}(0) = 0$ is of special interest; this occurs when

$$P_o = P_{crit} \equiv [8/(3(1-\nu)(5+\nu))] \cdot Et^3/R^2\rho\,.$$

At this critical pressure, the center of the wafer just touches the substrate. Now we find that the maximum gap occurs when $r = R/\sqrt{2}$, at which point

$$w_{gap}(R/\sqrt{2}) = 0.054 \cdot w_{sub}(0).$$

This means that when the suction pressure is barely sufficient to bring the silicon chip or wafer into contact with the substrate at the center (i.e., $P_o = P_{crit}$), the maximum gap is now only 5.4% of its original undeflected value. This maximum occurs approximately 71% from the center of the wafer.

Further increases in pressure would result in a growth of the central contact area. The analysis becomes quite involved at this point because the central support area is a function of load. We would expect that as the pressure is increased further, new contact areas would develop, resulting in an even smaller maximum gap between the surfaces. For example, an analysis of the analogous one-dimensional problem (using beam deflection theory) suggests that a further doubling of the pressure would reduce the minimum gap by another factor of 4, although the extendability of this result to two dimensions is questionable.

If we had tried flattening the silicon against a convex substrate, the results would have been less satisfactory. The maximum gap would occur at the edges of the die and approximately 10 times as much pressure would be required to achieve the same worst-case gap as for the

concave case, according to the author's calculations. For this reason, our heat sink substrates were deliberately designed with a slight concave curvature so that the physical situation resembles the case analyzed above, regardless of the sign of the wafer warpage.

4.3.3. Dust

In any practical fabrication sequence, some dust particles will be incorporated in the interfacial layer, creating increased interfacial gaps. The severity of this problem is determined by the fabrication environment. The critical portion of our fabrication sequence was performed in a "Class 100" clean room, which resulted in only an occasional trapped dust particle of sufficient size to degrade the thermal performance.

Since the density of detrimental dust particles can be kept low, we shall analyze the effect of a **single** particle between a perfectly planar substrate and a nominally planar silicon die. The surfaces are held together by a uniform suction pressure P_o (e.g., the capillary pressure). The trapped dust particle will create a gap in its vicinity. We assume that the effect of this particle is felt over some distance R; beyond that radius the gap is assumed to be zero (perfect contact). The particle is assumed to create a gap of distance w_d in its immediate vicinity. Note that w_d is not necessarily as large as the initial particle height, because some localized deformation of the particle and indentation of the plate may occur. w_d is the effective height of the dust, i.e., how high it locally lifts the wafer.

Fig. 4-10 is a sketch of our model for this situation, using elastic plate-deflection theory (w = vertical plate deflection).

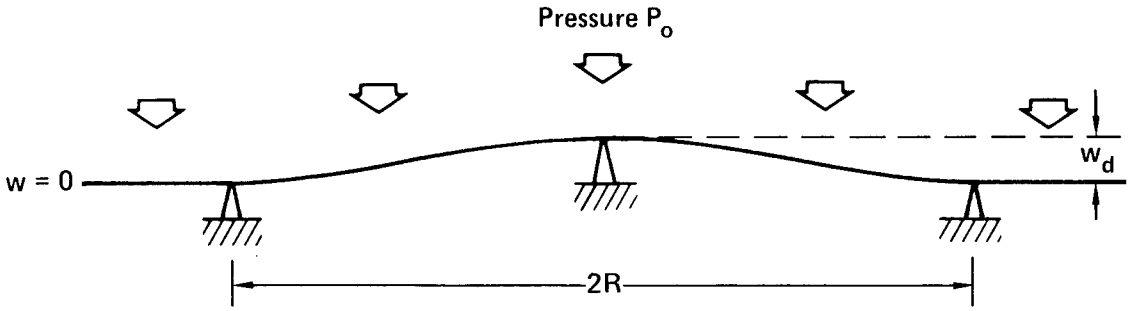


Figure 4-10: Elastic-plate model of a trapped dust particle.

The region of interest is a plate of radius R, centered at the dust particle, uniformly loaded with pressure P_o, and simply supported at the center to a height w_d. An additional simple

support at radius R is provided by the substrate (height w = 0). In addition we require that R be chosen to achieve tangency and osculation between the surfaces at r = R, i.e., $\partial w/\partial r = 0$ at r = R. For this radially symmetric problem, the general solution to Eq. (4.1) is:

$$w(r) = (ar^2 + b)\ln r + (cr^2 + d) - (P_o/64D)r^4 \tag{4.3}$$

The boundary conditions are:

$w(R) = 0$
$w'(R) = 0$
$w''(R) = 0$
$w(0) = w_d$
$w'(0) = 0$

The first three conditions imply simple support and parallel planes at r = R; the last two impose simple support at r = 0. The unknowns a,b,c,d, and R must have the following values to satisfy the boundary conditions:

$R = (64w_dD/P_o)^{1/4}$
$a = P_oR^2/16D$
$b = 0$
$c = -P_o(R^2\ln R)/16D$
$d = w_d$

Thus

$$w(r)/w_d = 1 + 4\rho^2\ln\rho - \rho^4, \text{ where } \rho = r/R\,. \tag{4.4}$$

This is plotted in Fig. 4-11. Note that a single micron-sized dust particle can elevate a relatively large area. For example, in Chapter 5 we tested a 100-μm thick silicon chip held by an 0.37-atm capillary suction force. If $w_d = 3\ \mu$m, a circular area of radius 3.0 mm is elevated to some extent; within 1.0 mm of the particle the gap is more than 50% of its maximum, i.e., w>1.5 μm. Since $D = Et^3/12(1-\nu^2)$, the most effective way to reduce the affected radius is to decrease the chip thickness t (i.e., make it more deflectable).

The effects of such a particle are mitigated to some extent by the compressive force on the particle. If no force were applied by the supports at r = R, the particle would be shouldering a force of πR^2P_o to hold the plate up. Actually only 25% of the total force is carried by the particle, i.e.,

$$F_{dust} = [2\pi rDw'(r)]_{r\to 0} = \pi R^2P_o/4\,.$$

The pressure on the dust particle will be the order of F_{dust}/A, where A is the load-bearing area of the dust particle. This will be very large because $A \ll \pi R^2$. For example, again consider a silicon chip having $t = 100\ \mu$m, $P_o = 0.35$ atm, and $w_d = 3.0\ \mu$m. We then have $F_{dust} = 2.47\times 10^4$ dynes. Assuming $A \simeq w_d^2/4$ (i.e., a roughly cubical dust particle), the particle would

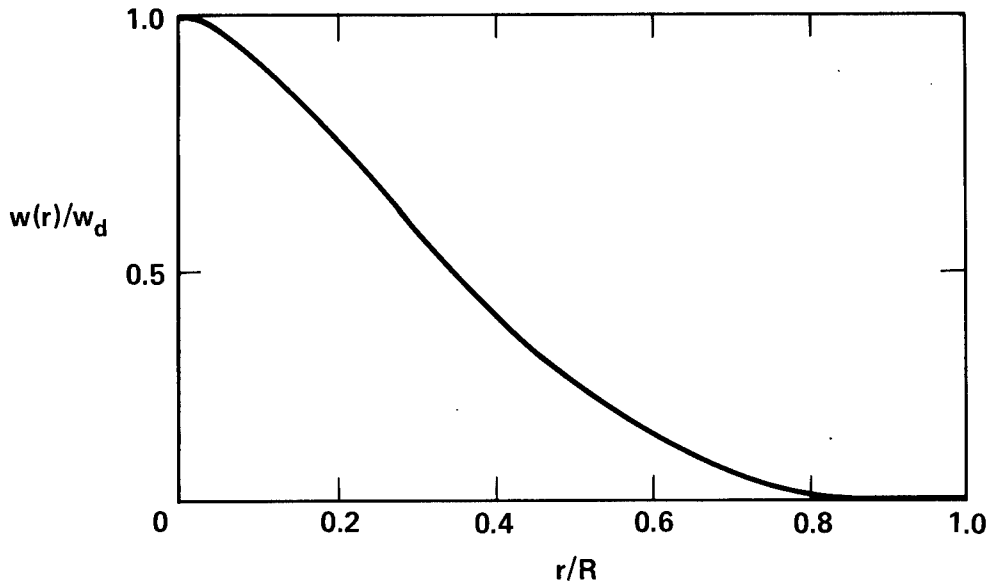


Figure 4-11: Predicted plate elevation around a trapped dust particle (Eq. (4.4)).

experience a compressive stress of about 1 Mbar. This is far larger than the yield stress for any material, and would typically correspond to a true strain of order unity. Thus the particle and/or wafer would plastically deform, flattening so that the contact area A increased until the stress was small enough to prevent further compressive deformation. The nickel-plated substrate (described in Chapter 5) would also deform, i.e., the particle would imbed itself in the nickel. The effective particle height would be greatly reduced by this deformation (perhaps by a factor of 2 or 3). Thus the effect of a dust particle is not as serious as might be inferred from its original dimensions; an isolated 3-μm dust particle might end up raising the silicon a maximum of only 1 μm! A smaller dust particle (e.g., 1 μm initial size) would be essentially completely squashed because of the greater pressure.

The preceding analysis has attempted to describe the effects of a few isolated dust particles. If a great many particles exist, their radii of influence R will overlap, and the analysis is no longer applicable. Instead the entire plate would be raised to a height roughly equal to w_d. We are not concerned with this case, because we fabricate the structures in reasonably clean environments.

4.4. Design

The basic design issues for microcapillary thermal interfaces (liquid selection and groove geometry) will now be described.

4.4.1. Choice of liquid

The choice of interfacial liquid is the most critical aspect of the design. An ideal interfacial liquid would have the following attributes:

1. High thermal conductivity.
2. High surface tension to maximize suction pressure.
3. A usable temperature range compatible with IC specifications (e.g., 0-70°C).
4. Very low vapor pressure (for longevity).
5. Good wetting characteristics (low contact angle), tolerant to slight contamination.
6. Newtonian rheology, with moderate or low viscosity (for ease of application and predictable capillary behavior).
7. Chemical compatibility with the mating surfaces.
8. Moderate cost.

Liquid metals might seem attractive, considering their high thermal conductivity and high surface tension. However, there are serious problems in meeting the other criteria which prevented their successful use in this work. There are only three candidates at room temperature: mercury (Hg), gallium alloys (e.g., Ga-In, Ga-In-Sn, or Ga-Sn-Zn), and various alloys of the alkali metals (K, Na, Cs) [95]. However, both Hg and the alkali metals are reactive, toxic, and very detrimental to semiconductor circuits. Having high-energy surfaces, liquid metals will only wet other high-energy surfaces (i.e., atomically clean metals) [96] with a suitably low contact angle. Even a monolayer of nonmetallic material such as oxide or organics will completely change the surface energy of a metal from high-energy to low-energy, making it nonwettable. Conceivably one could deposit an atomically clean layer of metal on the back of a silicon chip. However, when a liquid metal "wets" the clean metal surfaces it alloys to some extent, often forming a rather messy amalgam. The long-term integrity of the metal film would then be questionable. It is clear that the proper use of liquid metals would require careful metallurgical and reliability studies as well as scrupulous cleaning procedures. While we do not rule out the possibility, liquid metals were not successfully used in our experiments.

Silicone oils (polymethylsiloxanes) appear to be adequate in all respects for use as an interfacial liquid. Their room-temperature thermal conductivity is typically 1.5 mW/cm-K, hence a maximum gap of less than 1.5 μm would be required to achieve R<0.1 W/cm^2-K. This

is readily achievable, based on our calculations of wafer warpage and plate deflection. The surface tension of Dow-Corning 705 (a diffusion-pump silicone oil) is 36.5 ergs/cm^2 [97]. Silicones are economical, extremely inert and wholly compatible with integrated circuits. Most importantly, they have extremely good wetting properties, even on low-energy surfaces. Specifically, the contact angle is low (a few degrees) for all but the lowest-energy surfaces such as Teflon [96]. Furthermore the higher surface-tension silicone oils are "autophobic" [96], meaning that they coat high-energy solid surfaces in such a way as to reduce the surface energy, preventing the contact angle from becoming equal to zero. That is, most silicone oils are "nonspreading". In contrast, a liquid which has zero contact angle will spread out over anything it contacts, creeping around corners in order to wet the maximum possible area. This would seriously limit the lifetime of our microcapillary interfaces because the oil would eventually escape from the ends of the grooves by spreading. The value of silicones as additives to oils to make them nonspreading is well known. They have been used in watch bearings and in spacecraft, where it is necessary to have indefinite lubrication [98]. To verify this behavior, a small droplet of Dow Corning 704 oil ($\gamma = 37.3$ ergs/cm^2) was applied to a nickel-plated silicon wafer such as was used in the experiments of Chapter 5. The oil spread until the contact angle was only a few degrees, and then never spread any further over a period of 3 months.

Our only other concern is the evaporative lifetime of the interfacial liquid. This is extremely long for silicone oils, even in an open environment. The vapor pressure of Dow-Corning 705 at 25°C is 4×10^{-8} Pa (3×10^{-10} torr); this corresponds to an evaporation flux of $\sim10^{-11}$ gm/cm^2-min [99]. Only the ends of the channels are exposed to ambient (Fig. 4-2), so for our (2 cm)^2 die, the total effective area of evaporation was roughly 5×10^{-3} cm^2. Complete depletion of an initial 3-μm layer of oil would thus require 5×10^4 years, which is several orders of magnitude greater than that required in practical applications. One might think that at 70°C this lifetime would be reduced to only 50 years; however, this is determined by the temperature at the ends of the microcapillaries. Normally one would not design the heated area to extend the entire length of the chip; there would usually be a millimeter or so of unheated length at the chip edge. The temperature of this region would be comparable to that of the underlying cold plate, i.e., 25°C or so. Any oil vapor in the capillaries underneath the main heated chip area would quickly recondense in the cool areas, hence only the temperature (and vapor pressure) at the cool ends of the channels determines the oil lifetime. The hot vapor cannot escape out the ends because the mean free path length before hitting a cold wall is extremely short (a few μm in our 2-μm wide capillaries), and the enthalpy flux due

to latent heat is negligible at these very low mass fluxes. Note that our design of locating the capillaries in the substrate/cold plate (rather than in the chip) minimizes such vapor generation, because the only exposed menisci are in the cold plate, which is substantially cooler than the chip.

It should be understood that there are a number of design modifications which could be made to dispense with the open-endedness of the capillaries and hence reduce even further the evaporative losses, while retaining the stability and non-voiding characteristics of the interface. These may be attractive when using higher vapor-pressure liquids, but there is no evident need for these designs at present, so we have not pursued them in detail.

4.4.2. Capillary Dimensions

The dimensions of the capillary grooves are primarily determined by the expected minimum gap and by the need to accommodate slight variations in fabrication conditions. As discussed, we require the groove width at the nominal meniscus level to be greater than the maximum expected gap width. Wafer warpage calculations indicate that for a $(2\ cm)^2$ die, a maximum concave bow of 15 μm can be flattened to less than 0.75 μm by a pressure of 0.25 atm or more. Thus we designed for a nominal average capillary width of 2 μm, which is well in excess of the 0.75 μm minimum. To achieve sufficient reentrant taper, a nominal neck width $W_{min} = 1\ \mu m$ was chosen, tapering to $W_{max} = 4\ \mu m$ at the bottoms.

The depth of the capillaries is noncritical, but deeper is better because it allows more liquid to be used and hence makes the assembly more tolerant of fabrication variations and subsequent mechanical flexure. For the same reasons, the density of capillaries per chip should be as large as practical, while still leaving enough planar surface area to get low interfacial thermal resistance. We chose a capillary depth of 30 μm and a period of 10 μm for the experiments of Chapter 5.

Chapter 5
Microcapillary Interface: Experiments

5.1. Fabrication

5.1.1. Selection of a Fabrication Technique

As discussed, the ideal capillary structures for implementing the microcapillary thermal interface would be a two-dimensional reentrant array of high-aspect-ratio grooves, as sketched in Fig. 4-3b. Alternatively, a one-dimensional array of trenches having occasional tunnels (Fig. 4-3a) would work, and this was the configuration used in our work. The capillaries should be about 1 μm wide at the necks and wider further down. The top surfaces should be coated with metal or some other ductile stress-absorbing material to protect the brittle silicon substrate from fracture during assembly.

Such reentrant structures are difficult to fabricate, and a number of different techniques were explored. Initially it was thought that plasma etching could be used, since there have been reports that reentrant grooves can be so fabricated under carefully controlled conditions [100, 101]. However the details of the constituent gases and plasma conditions were not available and we were unable to duplicate these results. Moreover the techniques are probably not capable of simultaneously providing high aspect ratio and reentrant taper.

In general, etching techniques (wet chemical or dry plasma) are not well suited to fabricating narrow, high-aspect-ratio reentrant grooves. This is because the mass transport of etchant into a deep, narrow groove is driven by diffusion (except at very low ambient pressure), leading to substantial concentration gradients along the groove. Thus the etchant concentration will be reduced deep within the grooves, resulting in a reduced etch rate. In addition, the product species must be expelled by diffusion. This results in a pile-up of product deep within the grooves, which may also slow the local etch reaction rate. Most etching techniques will thus produce normally tapered grooves, i.e., narrower at the bottom than at the top, and the effect is most pronounced when the grooves are narrow and deep. Even the nearly perfect anisotropic etching of <110> silicon using KOH exhibits a very slight, almost imperceptible taper (Fig. 3-7).

These characteristics of diffusion-limited reactions can be turned to our advantage if a **deposition** process is used. Fig. 5-1b shows our approach, in which nominally vertical high-aspect-ratio grooves (fabricated by ODE of <110> silicon) of width W_{cap} and depth L_{cap} are subjected to an isotropic coating process. The same diffusion processes as described above will result in a lower deposition rate deeper in the grooves, thus producing the desired taper.

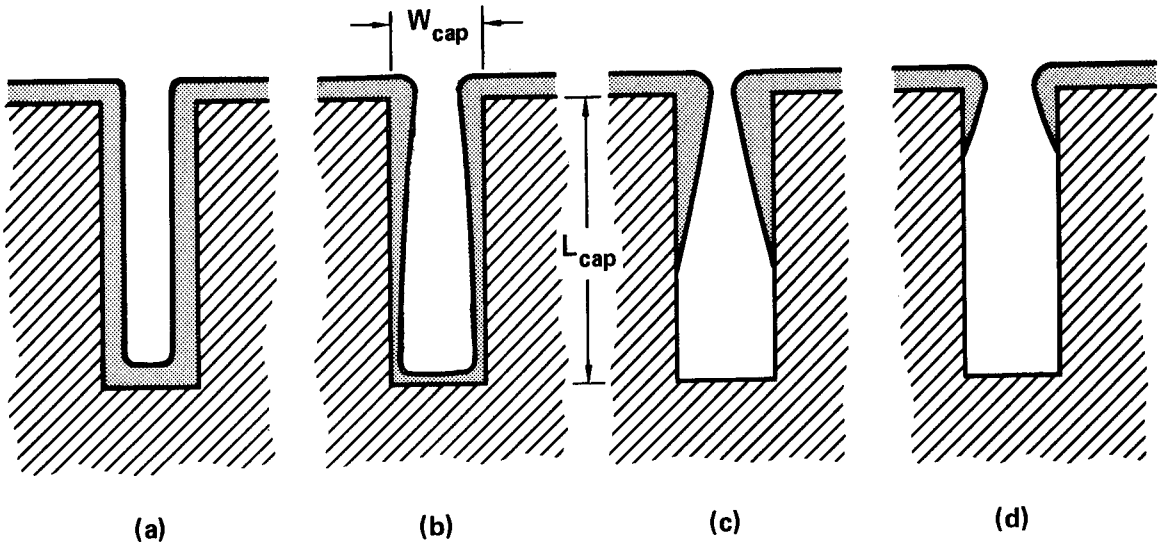


Figure 5-1: Isotropic deposition (e.g., CVD, oxidation, or electroless plating) under various kinetic conditions. In (a), the surface deposition is rate-limiting; in (d), diffusion into the groove is rate-limiting.

The degree of taper may be estimated by a simple argument. Let D be the diffusivity (cm^2/sec) of the deposition species in the grooves. The time taken for such species to diffuse over a distance x is approximately x^2/D. Now let τ_d be the time required for a monolayer of reactant to deposit onto the substrate. It is not usually necessary to know the complicated chemical kinetics which determine τ_d; one simply measures the deposition rate on an unobstructed surface.

We expect that if $\tau_d \gg L_{cap}^2/D$, then a fairly uniform deposition thickness will result over the entire groove, because the species have ample time to diffuse to the bottom (Fig. 5-1a). If $\tau_d \simeq L_{cap}^2/D$, we will expect a significant exponential taper in the deposited thickness, which is what we desire (Fig. 5-1b). (We are neglecting the fact that as the groove necks off, the deposition rate in the grooves will fall off even further.) If $W_{cap}^2/D \ll \tau_d \ll L_{cap}^2/D$, there will be an exponential taper in film thickness along the upper walls, and virtually no deposition

deep within the groove (Fig. 5-1c). This too is acceptable but not as desirable as far as fabrication tolerances are concerned; the lower portions of the grooves will not exhibit the desired capillary properties and hence the liquid quantity must be restricted so as to only fill the upper portions. Finally, if $\tau_d \ll W_{cap}^2/D$, there will be practically no taper in the groove, just an overhanging lip which is not sufficient (Fig. 5-1d) to give the desired stable capillary properties.

An example of the extreme deposition-rate limited case of Fig. 5-1a would be the thermal oxidation of silicon. For 1-μm thick oxide, the fastest practical oxidation rates are the order of 0.36 μm/hr for "wet" O_2 at 1200°C [102]. This corresponds to $\tau_d = 3.6$ sec. The diffusivity of H_2O in air at this temperature is approximately $D \simeq 0.5$ cm^2/sec. In grooves of depth 30 μm such as we are considering, $\tau_d \gg L_{cap}^2/D$ by more than 5 orders of magnitude, thus the grooves will oxidize uniformly.

An example of the other extreme (diffusion-rate limited deposition) would be most chemical vapor deposition (CVD) processes. In CVD depositions τ_d is very short (practically zero) because the species "stick" to the first surface encountered. Fig. 5-2 is an SEM of CVD-deposited SiO_2, which demonstrates this condition.

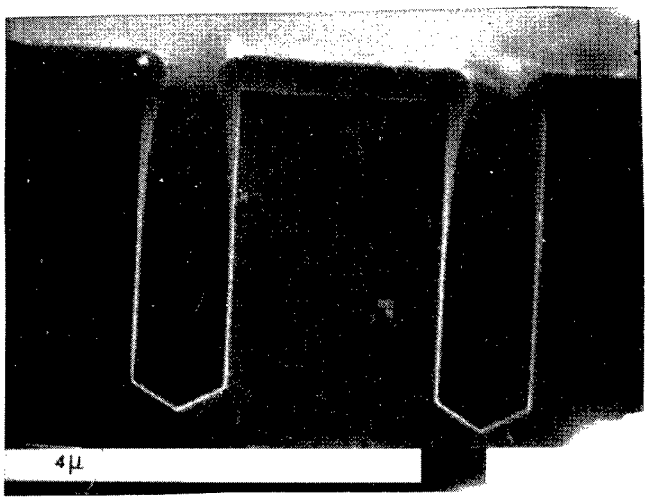


Figure 5-2: SEM of CVD-deposited SiO_2 on vertical silicon grooves.

It turns out that the desired case of $\tau_d \simeq L_{cap}^2/D$ may be achieved using an electroless metal plating process. The diffusivity of electrolytes in water at low concentrations is typically $D \simeq 1-3\times10^{-5}$ cm^2/sec [103]. Thus we desire $\tau_d \simeq (30\ \mu m)^2/D \simeq 0.5$ sec, which for a monolayer thickness of 0.55 nm corresponds to unobstructed deposition rates of 0.6 μm/min.

This is a typical electroless plating rate, as will be shown in the next section (Fig. 5-4). It is possible to adjust the reaction rate by varying plating conditions such as temperature, to obtain even more control over the groove shape.

An additional advantage of using of a metal-plating process is that the deposited film acts as a protective strain buffer, obviating the need for a separate metal deposition step to protect the silicon microcapillaries from fracture. For these reasons, electroless plating of anisotropically-etched silicon grooves was selected as the fabrication process for the microcapillaries. It should be noted that there are certainly other possible fabrication techniques and, in particular, abruptly-tapered grooves such as in Fig. 4-5 would also be acceptable; these could be fabricated by using a doped epitaxial layer and appropriate selective etches.

5.1.2. Electroless Nickel Plating

As discussed above, electroless plating of vertically-etched grooves was selected to fabricate the reentrant microcapillaries. Nickel was chosen because it has the most extensively characterized and most widely available electroless plating process [104]. A commercial plating solution (Anomet® 24 from M&T Chemicals Corp., Pico Rivera, Calif.) was selected. As with most electroless nickel processes, the mixed solution is an aqueous solution of sodium hypophosphite (NaH_2PO_2), a nickel salt, and proprietary catalysts. The sodium hypophosphite reduces the nickel cations under the action of the catalyst. Hydrogen is liberated during the reaction. The "nickel" (actually an amorphous nickel-phosphorus alloy containing 7-12% phosphorus) will plate out on suitably "activated" surfaces. The plating process is autocatalytic and hence once the plating has initiated somewhere it will continue until the solution is spent. For this reason great care must be taken to insure that the opportunity does not exist for nucleation of nickel deposits at any location other than the sample which is to be plated. Extraneous nucleation sites will result in the growth of spherical nickel particles which can then become imbedded in the smooth surface of the sample. Fig. 5-3 is an optical micrograph showing the results of such a condition.

Fortunately, activation of plating at normal plating temperatures ($T<90^{o}C$) will normally only occur at the surface of certain clean metals capable of catalyzing the reaction. The probability of activation can be enhanced by momentarily touching the sample with a steel or aluminum wire to initiate a galvanic action [105]. However, if the local solution temperature exceeds $90^{o}C$, plating could initiate spontaneously in the silicon, resulting in a rapid,

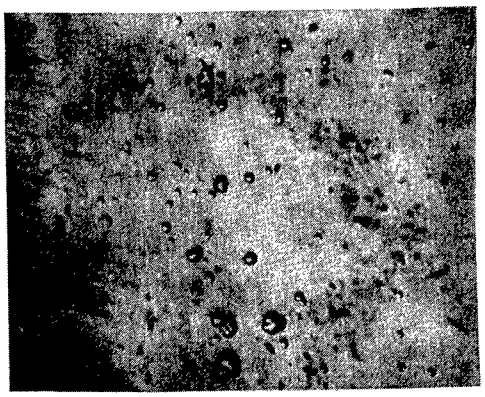


Figure 5-3: Hemispherical nickel particles due to impurities (nucleation sites) in electroless plating solution; magnification 700×.

catastrophic nickel precipitation within a few seconds. For this reason, the temperature of the plating bath must be held at a very uniform temperature. Localized heating with a burner is totally unacceptable; immersion of the beaker in a constant-temperature bath is preferred. For faster temperature cycling the solution may be heated in a clean glass beaker on a laboratory hot plate provided that a magnetic stirrer is used to minimize temperature variations to ± 1 or $2^{\circ}C$. When not in use, the solution temperature should be reduced to $70^{\circ}C$ or less to reduce the possibility of nickel precipitation. To further reduce this possibility, fresh solutions were frequently prepared. The beakers were precleaned first by immersion in a sulfuric-chromic acid mixture, rinsed, and then briefly filled with 1:1 HF-HNO_3, which had the dual effects of etching the glass slightly and dissolving any nickel-phosphorus particles which remained from previous plating sessions.

The most severe problem with electroless nickel plating was adhesion to the substrate. Activation is difficult and adhesion is very poor on lightly-doped silicon. Heavily-doped n-type silicon provides a more favorable surface for adhesion, and heat-treated electroless nickel has been used in the past to make ohmic contacts [106]. It was not convenient for us to dope the silicon grooves nor to anneal the deposited films. Instead we evaporated a thin ($\simeq$30 nm) layer of chromium onto the clean, bare silicon grooves and then, without breaking vacuum, evaporated 0.3 μm of nickel. Both the Ni and Cr evaporation sources were linear, longer than the sample width, and oriented perpendicular to the silicon microchannel direction. By choosing the filament height appropriately, complete coverage of the top, side walls, and bottom of each groove was assured. The chromium provided good adhesion to the bare silicon and to the deposited nickel, provided the source had been properly degassed prior to the evaporation. The nickel provided an easily activated surface for plating onto. The films

were allowed to cool in vacuum before being removed, thus preventing the formation of a native oxide on the nickel upon exposure to atmosphere. Even with this precaution, we sometimes observed a failure of the surface to activate uniformly in the plating solution. Attempts to force the plating to start by electrical bias resulted in films having poor adhesion and poor planarity (due to local nucleation at many points, rather than uniform plating over the entire surface). To obtain uniform activation, the sample was first dipped in a boiling NaH_2PO_2 solution to reduce any native nickel oxide. When bubbles appeared on the surface (presumably an indication of the alternate oxidation and reduction of the bare nickel layer), the sample was then immediately transferred to the plating solution. This procedure always resulted in immediate uniform initiation of plating, with resultant bright, smooth films. When these procedures were not followed, adhesion problems occurred.

Good adhesion is particularly important because the plated films are permanently under tensile stress and hence will peel when the thickness is sufficient to exceed the shear strength of the silicon/metal interface. The reasons for the built-in stress are complicated and the results highly dependent on solution composition [107]. We did not pursue the subject except to note that plated thicknesses in excess of 10 μm were required to cause peeling when the above procedures were followed. This was entirely adequate for our purposes, in which film thicknesses did not exceed 2 μm. Although thermal annealing in a forming-gas atmosphere can alleviate some of the stresses, very slow cooling ($\sim 5^{\circ}$C/min) is required to prevent cracking due to thermal stresses [108]. We therefore did not choose to anneal the nickel, which had adequate hardness and ductility as deposited.

The plating rate of the Anomet 24 was measured as a function of temperature in a series of experiments. Rectangles of silicon (precisely sawn to 1.000 cm $\times$ 5.000 cm) were evaporated with Cr/Ni and weighed in an analytical balance having microgram accuracy. They were then plated at various temperatures for a specified time and then reweighed. Several successive platings of the same sample at the same temperature were taken to eliminate possible errors related to the initiation of the plating process. Although a small amount of water was undoubtedly absorbed or adsorbed in the nickel-phosphorus films during plating, it was not practical to bake it out before weighing, because the sample would then slowly reabsorb 100-200 μg from the atmosphere, interfering with the weighing. Thus all samples were weighed in their full water-saturated state. In order to convert the weight measurements into thickness, knowledge of the density ρ of the nickel-phosphorus was needed. This was determined by measuring the thickness of a relatively thick (10 μm) layer of nickel using a

surface profilometer. We found that $\rho = 8.00 \pm 0.10$ gm/cm^3, which is less than that of pure crystalline nickel ($\rho_{Ni} = 8.9$) because of the phosphorus content and different microstructure.

Fig. 5-4 is a plot of measured plating rate as a function of temperature for a fresh batch of Anomet 24 in its recommended concentration.

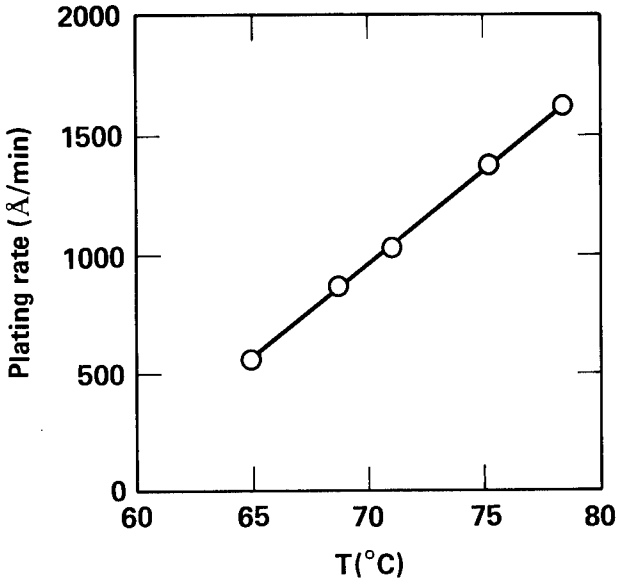


Figure 5-4: Measured plating rate vs. temperature of Anomet® 24 plating solution.

The plating rate will drop as the nickel is plated out from solution, falling to half its initial rate when approximately 0.27 cm^3 of nickel per liter of solution has been plated. For our purposes a nominal plating rate of approximately 0.1 μm/min was desired, so we plated at 70°C. Fig. 5-5 shows an SEM cross section of silicon microcapillaries which were successfully plated under these conditions. As predicted, the nickel film gradually thins down with increasing groove depth. This suggests that our approximate analysis in Section 5.1.1 is at least qualitatively correct.

5.1.3. Anomalous Behavior of Electroless Plating

We did not always obtain the nice reentrant structure of Fig. 5-5. Sometimes we would get a structure such as shown in Fig. 5-6. Here, the upper portions of the grooves are thickly plated as expected, but the plated film abruptly thins at some point part way down the groove. The position of this neck varies from groove to groove. This was rather puzzling because the plating conditions in both figures were thought to be identical. Inspection of Fig. 5-6 suggests

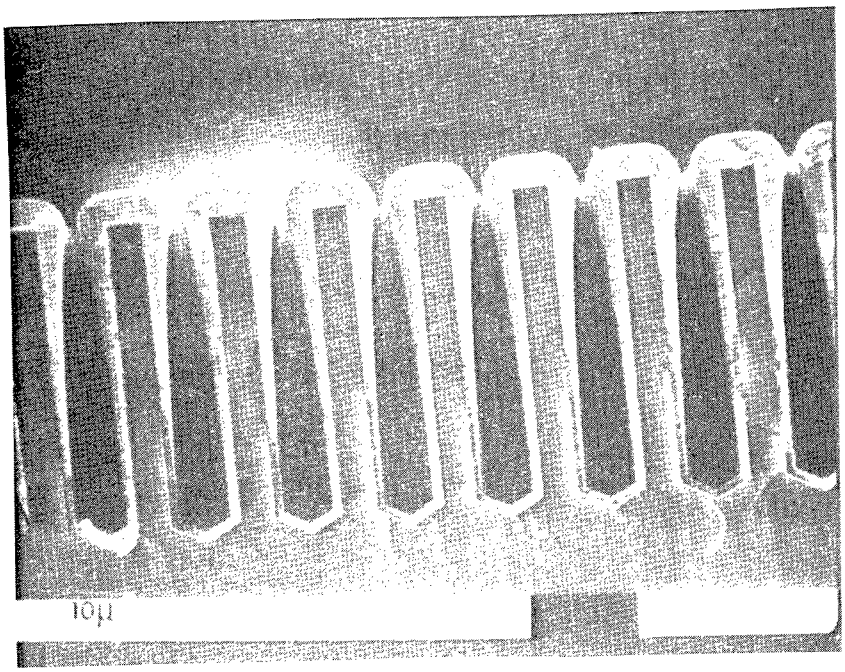


Figure 5-5: SEM of silicon grooves electrolessly plated with nickel (sample 82S29D2).

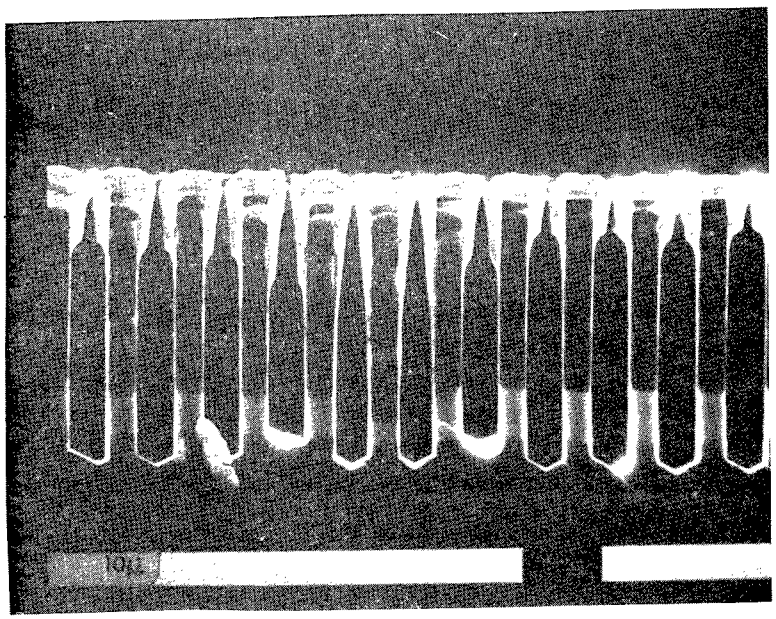


Figure 5-6: Anomalous shapes of electrolessly plated grooves, presumably due to trapped H_2 gas bubbles (sample 82D9B2).

a mechanism for producing such structures. We know that H_2 gas is evolved from the plating

process. Although initially an evolved H_2 molecule would be in solution at the moment of its creation (i.e., at the plating site), the concentration of dissolved H_2 will rapidly exceed its solubility in water (about 1 cm^3 of H_2 gas per gram of water) [109]. This causes a metastable condition, in which it is energetically favorable for the hydrogen to come out of solution but no easy pathway is available. Just as nucleation sites are required for a superheated liquid to begin boiling, so would nucleation sites be needed for the supersaturated hydrogen gas to come out of solution. Of course, once a bubble has formed it is free to grow until its vicinity is no longer supersaturated. We believe that in Fig. 5-6, H_2 bubbles nucleated in the grooves. This would not be a serious matter for normally tapered grooves, because in such grooves bubbles are not stable. In normally-tapered grooves, a trapped bubble will experience a net upward force due to surface tension and hence be expelled from the bottom of the groove (we assume zero contact angle, which is accurate for clean surfaces). In contrast, a bubble existing in a long **reentrant** groove will be flattened by surface tension, tending to remain at the bottom of the groove, while growing lengthwise. As more H_2 gas is generated in the vicinity, the bubble will grow longer and longer along the groove because that only requires overcoming modest viscous forces. The bubble will not grow vertically up the groove because this requires a substantial reduction in the meniscus radius as the bubble advances, which requires much larger pressures than are needed to grow lengthwise. The bubble will grow lengthwise until the bottom of the groove is filled, and its height will then be limited when the bubble's internal pressure is sufficiently high that other nucleation sites become more attractive to the dissolved hydrogen.

Inspection of Fig. 5-6 lends support to this "trapped bubble" hypothesis. The necking point is virtually circular, clearly suggesting that a bubble filled up the lower portion of the groove, blocking further plating. The exact position of the neck in each groove is variable, and is presumably determined by the time at which a bubble was nucleated in that particular groove. If the nucleation occurs late in the plating process, the surface tension forces which oppose bubble growth are stronger (due to the increased taper) and hence the necking point would be further down in the groove than if the bubble nucleation occurred earlier. Our model is further substantiated by our observation that for each groove, the neck point remained at the same depth along the entire length of the groove (confirmed by SEM examination of a series of cross sections). That is, the bubble grows by expanding lengthwise, this being the path of least resistance. Although we desired reentrant grooves because they stabilize the congregation of interfacial liquid at the necks and force voids deep down into the grooves (Fig. 4-2), this same effect ironically causes manufacturing difficulties during the electroless plating by trapping nucleated hydrogen bubbles in the grooves!

As far as the microcapillary thermal interface technology is concerned, the "trapped bubble" effect is not a critical problem. The grooves still have a reentrant shape, and the interface will have the required stability. However it was troubling that different results (Fig. 5-5 vs. Fig. 5-6) were obtained under nominally identical processing conditions. The hidden variable was finally determined to be the **age** of the solution. This suggests that the older solution contains relatively "easy" nucleation sites (sometimes called "motes" in the theory of boiling [110]) suspended in solution, which provide an attractive alternative to nucleation in the grooves. This is plausible because nucleation on perfectly smooth surfaces such as ours is known to be difficult, and hence would likely require a large supersaturation of H_2 to initiate bubble formation. In contrast, if easy nucleation sites (i.e., sites which don't require a large supersaturation of gas to nucleate) were floating around in solution, the supersaturation would remain low and nucleation would not occur in the grooves. If, on the other hand, no such "easy" sites are available, then nucleation would have to occur at a "difficult" site on the smooth substrate, and the bottom of the grooves are the best candidates because the H_2 supersaturation would be largest there, owing to the H_2 concentration gradient which exists due to diffusion out the tops of the grooves.

The nature of the mobile nucleation sites (motes) is not yet proven, but they may be microscopic nickel precipitates which plated out of solution due to random nucleation on the beaker walls, at hot spots, etc. These particles could be continuous sites for H_2 bubble nucleation as they meander throughout the solution by Brownian motion. In fact most of the time they would be connected to a growing bubble; when the bubble is sufficiently large its buoyancy would send it to the surface of the solution, where the mote would then detach and eventually migrate back to near the substrate to nucleate a new bubble. One objection to this model might be that the motes would also migrate into the grooves and start a trapped bubble there. Indeed there is evidence that this does happen to a small extent: a careful SEM examination of the sample 82S29D2 (from Fig. 5-5) revealed anomalous necking in about 5% of the grooves (not shown). Nonetheless, we believe that the the physics of the situation does not statistically favor trapping of motes. For a bubble to be trapped, a mote having either a very small associated bubble or none at all must diffuse into the groove and then not diffuse out again until it has grown too large to escape (diameter of 2 or 3 μm). But for a (e.g.) 0.1 μm-diameter nickel particle, Brownian motion imparts a velocity of $\sqrt{k_B T/m} \simeq 3$ cm/sec, i.e., only a one-millisecond residence time in the 30-μm groove. It can be calculated that the bubble diameter could not grow by more than 2.5 nm in this time. On the other hand if the mote is already associated with a bubble greater than 2 or 3 μm in size, it will be too large to

even enter the groove. Thus it seems plausible that the probability of a mote generating a trapped bubble is low, perhaps lower than the chance of spontaneous nucleation occurring if there are **not enough** motes around.

If our hypothesis about the formation of trapped bubbles is correct, then any modification which increases the availability of easy nucleation sites outside the grooves would greatly reduce the probability of having trapped bubbles. Fig. 5-7 shows the results of one such modification.

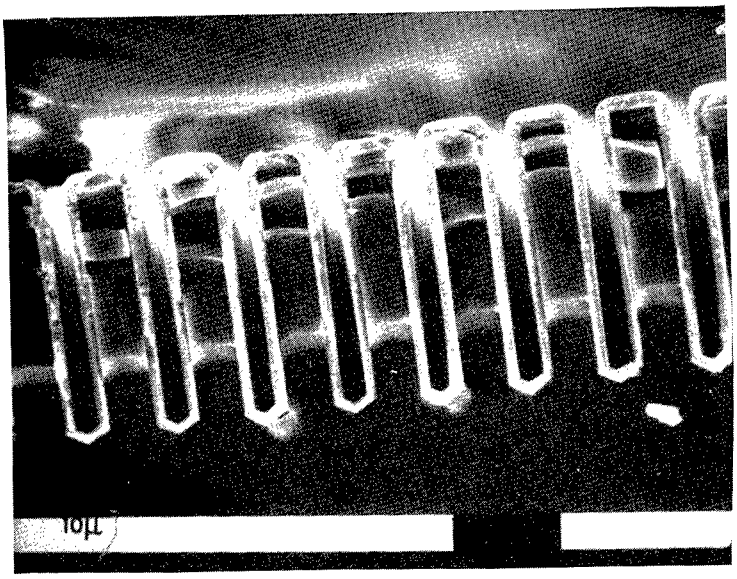


Figure 5-7: Silicon microcapillaries damaged (melted) with an E-beam prior to plating.

In this experiment, the silicon grooves were vertically etched and then Cr/Ni was evaporated onto them, as usual. Then the substrate was damaged with a scanning electron beam having just enough power density to momentarily melt the surface of the silicon. The tops of the fins balled up and, more importantly, dissolved the Cr/Ni coating so that the region would not get plated. The groves were then electrolessly plated in a clean, fresh solution (i.e., one favorable to bubble trapping). As the cross section in Fig. 5-7 shows, there is no evidence of trapped bubbles in the grooves; the nickel is quite thick at the bottom and tapers gradually towards the top until the damaged area is reached (contrast with Fig. 5-6). We conclude that the damaged portion was a good source of easy nucleation sites (particularly because it does not itself get coated with nickel); hence bubble trapping is suppressed.

Attempts were made to artificially introduce motes into solution by adding various fine

powders such as Ni, Ni_2O_3, or Al_2O_3. The results were inconclusive, however, and the powders tended to become embedded in the plated films, degrading the planarity of the surface. An interesting experiment which we have not yet tried would be to blow a jet of fine gas bubbles across the surface of the immersed sample as it is plated. The idea here is that these bubbles, having already been generated, would act as sinks for the evolved H_2, preventing nucleation of new bubbles.

For practical fabrication, we felt that the difficulties in trying to properly "age" the solution to eliminate the bubble-trapping effect were not worth the effort, so we used fresh, clean solutions. Although this led to bubble trapping, this does not interfere with the interface because the narrowest portion of the groove (at the top) is still accurately controllable, and this is the only critical dimension in the capillary fabrication.

5.1.4. Interface Fabrication Procedures

The most recent version of the processing sequence for the fabrication of a microcapillary thermal interface and combined micro-heat sink is detailed in Table 5-1 (refer to Table 3-2 for the description of certain subprocedures). The process at first parallels that of the plain micro-heat sink (Table 3-1). After the micro-heat sink grooves have been etched in the back of the silicon, the wafer is inverted and the tunnels (0.8 μm deep) are patterned and etched in the silicon. At this point the Pyrex/silicon anodic bond is performed on the back of the chip, with the bonding temperature chosen to create a concave curvature of the substrate. This is done to simulate wafer warpage, as discussed in Section 3.1.2. The curvature is verified by examining the substrate under an optical flat with monochromatic illumination. Typically the substrate bow was 15 μm from edge to center. The microcapillary grooves are vertically etched in KOH to a nominal depth of 30 μm, evaporated with Ni/Cr, and electrolessly plated with Ni. Extreme care is taken to insure cleanliness prior to and during the plating process, to prevent dust particles from becoming imbedded in the plated nickel. To avoid fracture, ultrasonic cleaning must not be used before the microcapillaries have been plated. Ultrasonic cleaning was found very effective for removing loosely bound particles **after** the plating is complete.

The plated grooves are carefully inspected for embedded particles using an optical microscope with grazing illumination (so that asperities cast a long, visible shadow). Despite our efforts to maintain cleanliness, small asperities ($\simeq$2 μm) were occasionally found. These could be flattened by laying the plated side flat against a clean, optically flat stainless steel

plate and applying mechanical pressure (30 psi). As discussed in Section 4.3.3, this will exceed the elastic limit of the nickel underneath the asperities and the asperities will therefore be driven into the nickel film.

The sample is then ready for capillary bonding to an integrated circuit chip. In our experiments a $(2\ cm)^2$ die having a $(1\ cm)^2$ WSi_2 thin-film resistor was used. The process sequence for fabricating the resistor chip is listed in Table 5-2, and is very similar to that used previously in the integral micro-heat sink fabrication (Section 3.1.5). A layer of Dow Corning 704 silicone oil is spun on the back of the heater resistor chip, and the chip is then immediately attached to the microcapillary sample. Typically a 3-μm layer of oil is used. The thickness t_{oil} is easily determined by the formula $t_{oil} = 3\nu/4\omega^2 t$ (from Section 3.1.2), where the kinematic viscosity of the oil was determined to be $\nu = 0.31 \pm .03$ poise. This formula was verified by ellipsometric measurements of the oil film thickness (measured index of refraction $\eta = 1.48 \pm 1\%$).

Since the heat sink substrates had a deliberate concave bow of $\sim$15 μm and the initial oil film thickness on the chip was only 3 μm, momentary pressure was applied to press the surfaces together. Once this was done, the surfaces remained in contact indefinitely, owing to the capillary attraction.

Table 5-1: Fabrication schedule for microcapillary thermal interfaces

Step	Process
1. Starting material	2", <110>, 500-μm thick, double-polished, lightly-doped (>1 Ω-cm) Measure flatness with optical flat Label flattest side as "front"
2. Oxidation	**RCA clean** 1100°C, 5'dry, 35'wet, 5'dry (0.5 μm nom.), front side up
3. Splay pattern lithography	**NPR photolith** (front side) Buffered HF etch, DI rinse 5', Nanospec inspect **DEWAX, STRIP** Inspect SiO_2 for pinholes or scratches
4. Splay pattern etch	50-50 KOH @ 70°C for 2 hrs (nom. depth 94 μm) Long DI rinse Identify <111> plane angle
5. Dice (for 2-sided alignment)	Back-surface protect (Nitto tape) Dice to $(34.5\ mm)^2$ (perfect square) Label edge $\perp$ to <111> plane (back side) Ultrasonic clean in soap, then H_2O_2/H_2SO_4 Rinse, switch to clean beaker Strip SiO_2 in 10:1 HF (ultrasonic agitation), rinse
6. Tunnel lithography	Singe 1 min. Back-surface protect (neg. photoresist) **NPR photolith** (front side) tunnel pattern
7. Tunnel etch	Singe 1 min. 50:1 HF squirt, blow dry Plasma etch 0.7 μ deep (20" SF_6, 500 W, 150 μ, 50 SCCM) **STRIP**
8. Oxidation	1200°C, 5' dry, 90' wet, 5' dry (front side up) slow push/pull! (t_{ox} = 1.1 μm nom.)
9. Microchannel lithography	**NPR photolith** (back), aligned w/<111> plane Front-surface protect Buffered HF etch, DI rinse 5', Nanospec inspect **DEWAX, STRIP**, inspect front SiO_2 for pinholes
10. Microchannel etch	Fresh, clean, 50-50 KOH 52°C (tilted) Etch 34 hrs (check @27 hrs); nom. depth 400 μm Long DI rinse

continued

Step	Process
11. Oxidation	**RCA clean** w/10:1 HF strip after H_2O_2/H_2SO_4 1000°C, 5' dry, 30' wet, 5'dry (front side up) very slow push/pull! (t_{ox} = 0.3 μm nom.)
12. Microcapillary lithography	Spin positive resist @5000RPM for 25 sec 15' prebake @95°C, 10' cool 40-45" exposure (clean mask) 42" develop, 20" rinse, inspect $\geq$10' postbake @90°C O_2 plasma descum 60", 500 Watts, 100 SCCM Buffered HF etch, thorough DI rinse, inspect H_2SO_4/Chromic strip, rinse
13. Prepare for anodic bonding	Front-surface protect (wax) Strip SiO_2 from back with buffered HF **DEWAX, STRIP**
14. Cover plate fabrication	Start w/optically polished Pyrex 0.8 mm thick Saw to (34.5 mm)$\times$(34.5 mm) (resinoid blade) Cut headers (sandblast or ultrasonic drill) Ultrasonic clean (solvent sequence) H_2O_2/H_2SO_4 clean
15. Cover plate bond	Anodic bond @ 400°C Measure laminate curvature with optical flat Saw to final size (25.4 mm$\times$34.5 mm)
16. Microcapillary etch	Ultrasonic clean in TCE Back-surface protect (seal headers w/tape, wax) 50:1 HF dip, dry 50-50 KOH @52°C for 2 hrs (nom. depth 30 μm) Rinse, strip SiO_2 in buffered HF **DEWAX**
17. Nickel coating	H_2SO_4/Chromic, rinse, methanol boil, oven dry 50:1 HF squirt, evaporate up Cr/Ni (4" filament height) Cool $\geq$30 min. in vacuum NaH_2PO_4 dip (30 sec. @90°C) Electroless plate (ultraclean, 70°C @15 min.) Rinse, TCE ultrasonic clean, TCE boil Ultrasonic clean (solvent sequence)
18. Bonding	Inspect, test flatness, flatten Resistor front-surface protect Spin interfacial oil onto resistor (3 μm nom.), attach resistor Package substrate (epoxy to Lexan®)

Table 5-2: Fabrication schedule for heater resistor

Step	Process
1. Starting material	100-μm thick, preferably double-polished Measure flatness with optical flat Label flattest side as "back"
2. Oxidation	**RCA clean** 1100°C, 5' dry, 35' wet, 5' dry (0.5 μm nom.)
3. Resistor deposition	Sputter WSi_2, 65' (2.34 μm nom.) Sputter SiO_2 (0.2 μm) for etch mask
4. Resistor lithography	**NPR photolith** (1 cm $\times$ 2 cm rectangle) Back-surface protect (seal headers) Buffered HF etch **DEWAX, STRIP** Etch WSi_2 in 1:3:4 HF:HNO_3:HAc ($\sim$75 sec), continue until SiO_2 mask gone ($\sim$60 sec more)
5. Contact metallization	**NPR photolith** (contact rectangles) Evaporate Cr/Ag (1 μm) Lift-off in hot, dry ECOSTRIP (ultrasonic agitation) Saw to (2 cm $\times$ 2 cm), ultrasonic solvent clean Solvent clean (TCE/Ace/Meth boil)
6. Bond contacts	Solder Au-clad Mo rectangles using indium foil Measure resistance
7. Paint for IR microscopy	Protect contacts Use "flat black" (high emissivity) spray paint

5.2. Experiments

5.2.1. Measurement Techniques

The thermal properties of microcapillary thermal interfaces were measured primarily by infrared microscopy. Movable thermocouple probes could not be used because such probes exert a localized force on the resistor, pressing the surfaces together locally and hence reducing the thermal resistance below its normal value. In contrast, the infrared microscope does not contact the silicon and hence does not mechanically perturb the interface.

The heater resistor sample was spray-painted with a thin, flat black paint to yield a uniform, high-emissivity ($\varepsilon \simeq 0.95$) surface. The IR microscope was a Barnes Model 2A [111] having a liquid nitrogen-cooled InSb infrared detector. It is capable of both dc and thermal transient measurements, the latter mode having 8 μsec response. A 36$\times$ reflective objective was used, which provides an 18-μm diameter spot size. This is a much smaller dimension than the thickness of the resistor's silicon substrate, hence this may be considered an accurate measure of local temperature. The IR microscope was calibrated by passing preheated water (from an upstream silicon heat sink) of known temperature through an unpowered test heat sink, measuring the surface temperature, and comparing with that measured by a thermocouple probe as in Section 3.2.2. The IR microscope has a claimed measurement accuracy of $\pm 1^{\circ}$C. A temperature-sensitive lacquer having a calibrated melting point of $74 \pm 1^{\circ}$C was also used; the two calibration techniques agreed within 1°C. Temperature maps were taken by manually scanning the sample with precision micropositioners.

5.2.2. Thermal Resistance Maps

Several microcapillary thermal interface samples were thermally mapped using the infrared microscope. Dow Corning 704 silicone oil was used as the interfacial liquid. The substrates had concave bows of about 16 μm, by design. The chips typically had dimensions (2 cm)$\times$(2 cm)$\times$(100 μm). As in the previous experiments, a (1 cm)$\times$(1 cm) WSi_2 thin-film heater resistor was used to supply power to the front surface of the chip. The infrared microscope measures the chip surface temperature $T_s(x,y)$ as a function of position. By dividing by the supplied power density $\dot{q}''$ (W/cm^2), we get the <u>total</u> normalized thermal resistance $R_{tot} = R_{chip} + R_{int} + R_{heat\text{-}sink}$. We could then deduce the interfacial thermal resistance R_{int} by subtracting the calculated values of R_{chip} and $R_{heat\text{-}sink}$, although of course there is some uncertainty in this procedure when R_{int} is small relative to the other thermal resistances.

In general R_{int} was in fact found to be quite low (typically $R_{int} \simeq .02$ cm^2-°C/W, which is about 20% of R_{tot}). Thus, the microcapillary thermal interface technology can indeed provide a thermal contact which does not greatly increase the overall thermal resistance of the resistor-to-ambient heat path. Table 5-3 is a typical 2-dimensional thermal map of R_{tot} for a carefully fabricated interface at a power flux of 242 W/cm^2.

x / y	1 mm (flow →)	2	3	4	5	6	7	8	9
1 mm	.088	.085	.096	.100	.101	.103	.107	.106	.100
2	.070	.082	.103	.109	.107	.103	.101	.103	.099
3	.057	.074	.095	.108	.109	.109	.109	.111	.108
4	.063	.080	.096	.107	.109	.111	.113	.114	.109
5	.068	.080	.094	.108	.111	.110	.110	.106	.101
6	.059	.068	.084	.105	.111	.113	.112	.106	.096
7	.046	.055	.076	.095	.103	.109	.111	.106	.093
8	.044	.057	.076	.089	.095	.102	.108	.107	.093
9	.042	.058	.076	.082	.085	.092	.102	.102	.089

R_{tot} (cm^2-°C/W)

Table 5-3: Thermal resistance (R_{tot}) map of a typical chip/interface/heat sink assembly.

The peak value of R_{tot} is .114 cm^2-°C/W, occurring at (x = 8 mm, y = 4 mm). At this point we calculate $R_{chip} \simeq .007$ and $R_{heat\text{-}sink} \simeq .085$; thus $R_{int} \simeq .022$ cm^2 – °C/W, which is quite small. This is a typical value; the worst-case value occurs at (x = 1 mm, y = 1 mm), where R_{int} is estimated to be .046 cm^2-°C/W. This is still well within our design goal of $R_{int} < 0.1$ cm^2-°C/W. Estimating the thermal conductivity of the oil at 1.5 mW/cm-K (the exact value has not been measured for DC-704 but this value is typical of silicone oils), we conclude that the apparent gap thickness is typically 0.33 μm and never more than 0.7 μm. Thus, despite the built-in substrate bow of 16 μm, the capillary suction has demonstrated its ability to hold the chip in conformance with the substrate to better than 0.7 μm!

The thermal resistance of our samples increased with increasing power flux. Since both the silicon chip and the silicon heat sink have fairly linear thermal resistances, this

nonlinearity is attributed primarily to a decrease in the thermal conductivity k_{704} of the DC-704 silicone oil with increasing temperature. The lower curve in Fig. 5-8 plots R_{int} (the calculated heat sink and chip thermal resistances have been subtracted out) for a typical point on one sample (not as good as the sample in Table 5-3) as a function of surface temperature. Here surface temperature was set by adjusting the power flux; the input water temperature was held constant at 19°C). R_{int} increased threefold as the power flux was increased to yield a 86°C chip temperature.

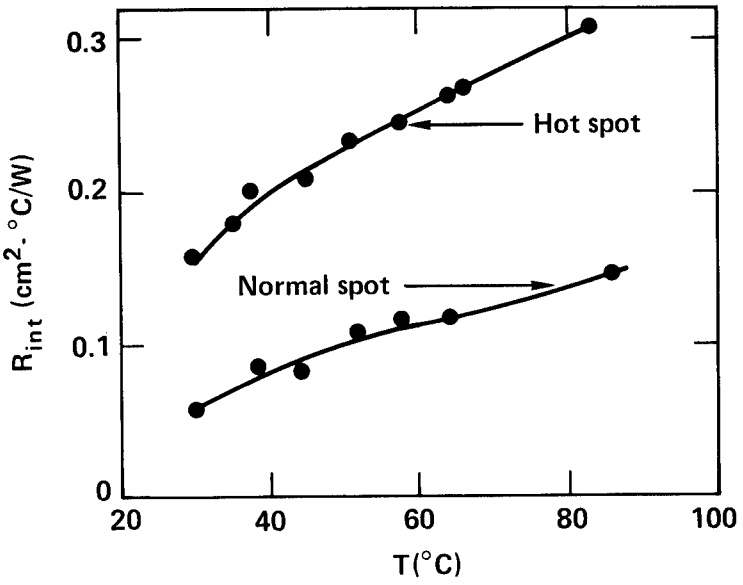


Figure 5-8: Thermal resistance vs. temperature of a typical spot and a hot spot.

5.2.3. Hot Spots

The quality (i.e., smallness of R_{int}) of our thermal interfaces was greatly influenced by environmental cleanliness during fabrication. As explained in Section 4.3.3, dust particles are expected to have a detrimental effect on the interface, locally increasing the gap thickness and hence the thermal resistance. We observed such behavior (local "hot spots") on samples which had not been as carefully processed. The upper curve in Fig. 5-8 plots R_{int} at the peak of a hot spot. As expected, the thermal resistance is nonlinear due to the temperature dependence of k_{704}. The curve is not an exact multiple of the lower curve, presumably because thermal spreading in the silicon allows some heat to bypass the high-resistance hot spot as it becomes more and more thermally resistive. The thermal resistance values suggest that a 3-μm gap exists at the center of the hot spot.

Two-dimensional maps were made of the temperature in the vicinity of the aforementioned hot spot. As expected, the hot spot exhibited nearly circular symmetry (no θ-dependence). Using these data and the thermal resistance data of Fig. 5-8, and estimating $k_{704} \simeq 1.5$ mW/cm-K at 60°C, we can deduce gap width as a function of distance from the hot spot, as plotted in Fig. 5-9. Using the theory of Section 4.3.3, we have also plotted the predicted gap due to an asperity of the same height.

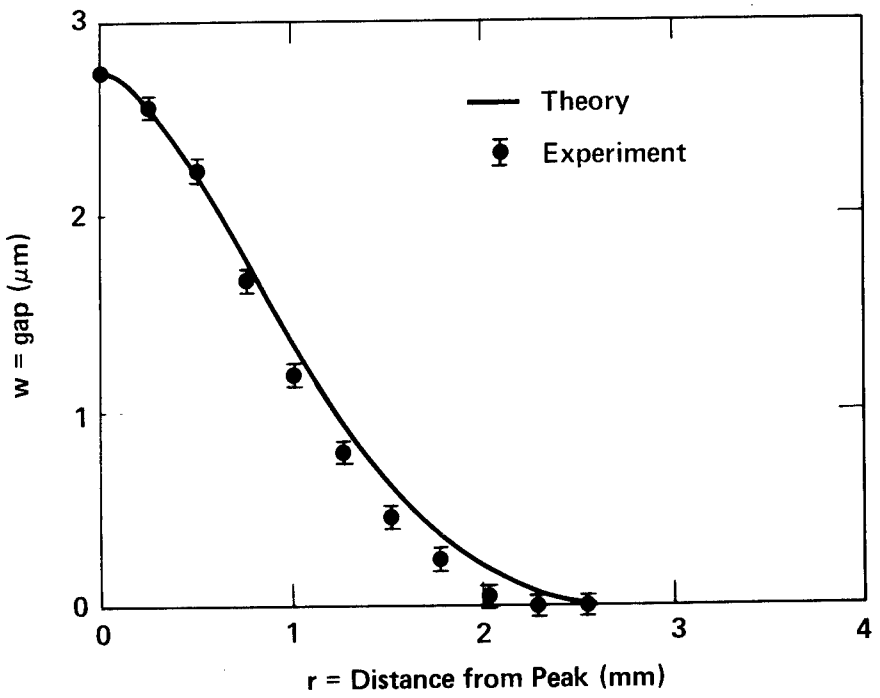


Figure 5-9: Interfacial gap near a hot spot (predicted and experimentally deduced).

The agreement is quite good in view of the simplicity of the model.

The origin of this hot spot was later explained when we detached the surfaces and inspected them under an optical microscope. The area of the nickel-plated substrate where the hot spot was measured was observed to contain a cluster of visible bumps several microns high, similar to those shown in Fig. 5-3. Evidently some particulate contamination occurred prior to or during the electroless plating process. We confirmed that the hot spot was caused by bumps on the substrate by reattaching the chip to the substrate at an offset location. The hot spot was found to have moved relative to the chip, but not relative to the substrate.

We conclude that asperities several microns tall can significantly degrade the performance

of the microcapillary thermal interface, as expected. However, we only observed the problem when an area was greatly contaminated. We did not observe any significant hot spots associated with single dust particles. Whether this was due to luck (i.e., no large dust particles were incorporated during assembly) or whether the structure is much less sensitive to single particles than the infinitely-strong dust-particle theory of Section 4.3.3 predicts (e.g., due to plastic deformation of the dust), we cannot yet say. But the main point is that with proper cleaning procedures and fabrication in a clean room, we were able to get very good thermal performance, as the example in Table 5-3 shows.

5.2.4. Long-Term Reliability

We expect the microcapillary thermal interface technology to be very reliable because of its freedom from shear and in-plane tensile stresses in the interface layer. In particular we expect virtual immunity to fatigue from repeated thermal cycling, because there is no solid adhesive or solder layer undergoing repeated strain. The only wear would be due to sliding of the surfaces due to relative expansion mismatches, which for smooth surfaces is very small [72]. To check this, we thermally cycled an interface assembly between 20°C and 120°C for 2×10^6 cycles at a rate of 3 cycles per second. (This was feasible because the thermal time constant of the structure is less than 10 msec.) As shown in Table 5-4, the total thermal resistance R_{tot} dropped significantly, typically by 20 to 30%; this corresponds to approximately a 50% drop in the interfacial thermal resistance R_{int}! This is attributed to the smoothing of surface asperities (trapped dust) caused by the sliding. This behavior contrasts with conventional die attach techniques, in which the thermal resistance of the bond invariably degrades with repeated thermal cycling as voids and fatigue cracks develop in the solder layer. (In fact, a measurable increase in thermal resistance is the first indication of an impending failure in conventional die attachments [76].) Subsequent disassembly and inspection of the microcapillary interface revealed no visible effects as a result of the thermal cycling, adverse or otherwise. Thus the attachment appears very reliable with respect to thermal cycling.

Since the microcapillary thermal interface does not make a permanent metallurgical bond, there may be questions as to the lateral mechanical stability of the chip. That is, will the chip move horizontally in response to thermal cycling, shock, or vibration? We observed no net motion of the substrate after 2×10^6 thermal cycles, within the measurement error of ±1 μm. To estimate the consequences of shock or vibration, the shear strength of the attachment was measured to be 9.4×10^4 dynes (95 ± 5 gm). Since the suction force is estimated to be (2

x / y	flow →											
	.036	.064	.110	.119	.106	.102	.114	.123	.117	.153	.282	.242
	.049	.074	.114	.117	.110	.114	.136	.157	.144	.159	.248	.231
	.057	.081	.108	.108	.108	.119	.142	.172	.155	.144	.157	.144
	.064	.089	.114	.114	.117	.125	.138	.159	.165	.148	.131	.089
	.061	.089	.112	.112	.112	.121	.127	.146	.146	.142	.123	.072
	.055	.078	.100	.104	.106	.114	.119	.123	.119	.114	.102	.055
	.044	.068	.091	.100	.102	.110	.112	.106	.098	.091	.081	.038
	.040	.064	.091	.100	.104	.106	.102	.095	.083	.078	.076	.042
	.040	.066	.093	.098	.100	.098	.098	.093	.081	.072	.068	.036
	.042	.059	.076	.078	.078	.080	.083	.083	.074	.061	.053	.030

R_{tot} (cm^2-°C/W)

(before cycling)

x / y	flow →											
	.039	.050	.095	.103	.095	.086	.097	.105	.097	.118	.217	.234
	.046	.059	.097	.103	.095	.097	.111	.131	.120	.124	.188	.217
	.052	.065	.092	.095	.095	.099	.116	.143	.135	.114	.120	.120
	.056	.073	.097	.099	.101	.105	.114	.128	.135	.120	.103	.078
	.054	.073	.097	.099	.099	.101	.107	.114	.116	.111	.099	.065
	.050	.065	.086	.090	.095	.097	.101	.101	.097	.090	.082	.052
	.044	.056	.080	.088	.092	.092	.097	.088	.078	.073	.067	.039
	.042	.054	.080	.099	.092	.090	.088	.080	.069	.063	.061	.039
	.039	.054	.080	.086	.088	.084	.084	.080	.069	.059	.054	.037
	.042	.052	.069	.071	.071	.073	.071	.071	.063	.052	.044	.029

R_{tot} (cm^2-°C/W)

(After $2x10^6$ thermal cycles)

Table 5-4: Thermal resistance maps ($R_{tot}(x,y)$) before and after 2 million thermal cycles.

$cm)^2 \times (0.37\ atm) = 1.48 \times 10^6$ dynes, this corresponds to coefficient of friction $\mu \simeq 0.06$, which is consistent with that expected between smooth, lubricated surfaces. Since the wafer only weighs .093 gm, a horizontal shock or vibration acceleration of $10^3 \cdot g$ (g = gravitational acceleration) would be required to slide the chip against friction. This is far in excess of that likely in commercial environments, so we expect the interface to be mechanically stable. If one is still concerned about sliding, a pair of locating surfaces might be microfabricated to prevent sliding while still permitting unrestricted differential thermal expansion of the surfaces.

Chapter 6
Summary and Conclusions

6.1. Results and Contributions

The main results and contributions of this work are:

- The fundamental limits involved in compact heat transfer from integrated circuits have been critically examined. Contrary to published claims of a 20 W/cm^2 heat flux limit for cooling densely packed arrays of ICs, it was found that by optimizing the convective heat transfer and by eliminating unnecessary thermal interfaces (i.e., constructing an integral heat sink), no fundamental or technological limit prevents the **compact** removal of well over 1 kW/cm^2 from 1-cm^2 ICs. This contradicts the widely held belief that certain high-performance logic families such as ECL are unsuitable for VLSI system applications due to their high power consumption.

- An optimization procedure was developed which allows easy calculation of the optimum dimensions for a liquid-cooled laminar flow heat sink. The duct and fin dimensions are microscopic (typically 50 μm), high-aspect-ratio (typically 8:1) structures. For a $(1\text{-cm})^2$ water-cooled silicon IC, thermal resistances of as low as .06 cm^2-°C/W are predicted at ordinary pressure drops (50 psi).

- Although traditionally turbulent flow is used to obtain high heat transfer, for a given pumping power expenditure a laminar-flow design becomes preferable as heat exchanger channel lengths are scaled down. The scaling could in principle be continued to allow an IC to operate continuously at 50 kW/cm^2 while maintaining its temperature below 120°C; multiple headers and high pressures would be needed to achieve such an extreme performance level.

- IC microfabrication techniques such as microlithography, orientation-dependent etching, and anodic bonding were adapted to fabricate these new water-cooled microscopic silicon heat sinks, having dimensions an order of magnitude smaller than in conventional compact heat exchangers. The microfabrication techniques provided great manufacturing precision (submicron tolerances), which was necessary in order to achieve optimal heat transfer.

- A test apparatus was developed to precisely measure silicon heat sink thermal and flow-friction performance. Very good agreement was found with theory. A peak thermal resistance of .083°C/W was measured for a $(1\text{-cm})^2$ heated silicon area, which allowed 1300 W/cm^2 to be continuously dissipated while maintaining a surface temperature of less than 130°C. This is more than 20 times lower than in state-of-the-art commercial technologies for cooling arrays of ICs; the latter

exhibit thermal resistances of $\sim$2 cm^2-°C/W [13, 14, 15]. If one uses power dissipated per unit volume as a figure of merit, then this technology can dissipate 10 kW/cm^3. This is more than four orders of magnitude greater than is generated in state-of-the-art computers, which typically generate at most 0.5 W/cm^3. Hence it can safely be concluded that thermal considerations need not pose any limit to the design of much denser and correspondingly more powerful computer systems.

- The high-performance heat sinking technologies developed in this work can also be used to enhance the reliability of ICs, because most IC failure mechanisms are very sensitive to temperature. For example, the rate of electromigration [1] in a conductor is proportional to $J^n \cdot \exp(-E_a/k_B T)$, where J is the current density and E_a is an activation energy (typically the order of 1/2 eV). By maintaining a high-power VLSI IC at 20°C, rather than 120°C, the median time to failure could be increased by 1 or 2 orders of magnitude.

- When properly packaged, silicon heat sinks exhibited high reliability. No degradation was measurable after 1000 hours of full-flow operation, nor was any expected.

- This use of an integrated silicon heat sink to cool high-performance ICs is a powerful solution to the cooling problem. If, however, the integral configuration is deemed impractical, a novel die attachment technique has been developed in which the surface tension of a liquid, partially filling an array of reentrant microcapillaries, holds the polished back of such an IC in intimate thermal contact with the heat sink. The main advantages of this technique over conventional die attachments using solders are: complete immunity to thermal expansion mismatch, inherent freedom from voiding, and unlimited ability to detach and reattach chips to the heat sink. It also provides a solution to the classic problem of heat-sinking and vacuum-chucking wafers in a high-vacuum environment. As such it is likely to be particularly useful for the testing and packaging of "wafer-scale" integrated circuits, where the aforementioned problems are particularly severe.

- IC-industry techniques such as microlithography, orientation-dependent etching, and electroless nickel plating were adapted to successfully fabricate microcapillary thermal interfaces. Measured interfacial thermal resistances were well within design limits ($R_{int} \simeq .022$ cm^2-°C/W typical; $R_{int} < .046$ cm^2-°C/W maximum). The thermal performance **improved** after 2×10^6 thermal cycles, in contrast with conventional die attachments. Successful detachment and reattachment of the interface was readily demonstrated. The effects of entrapped dust particles larger than 3 microns are detrimental to the performance of our designs, but fabrication in a standard IC Class 100 clean room eliminated this problem.

6.2. Recommendations

6.2.1. Thermal Considerations

The following are recommendations for further work in enhancing compact heat transfer from ICs.

- Investigation of the limits to which heat transfer can be increased by scaling down channel lengths and incorporating multiple headers in silicon heat sinks. This could yield thermal resistances of as low as .002 cm^2-°C/W, but this has not been proven. This topic is not relevant to IC applications because the present demonstrated thermal resistance of .083 cm^2-°C/W is far in excess of current or foreseeable needs; however it is an interesting research area for other heat-transfer applications.

- Investigation of micro-heat sink behavior at cryogenic temperatures, where the thermal conductivity of silicon is extremely high and phonon mean free path lengths are comparable to fin dimensions. This would be important if this technology were to be applied to cooling of cryogenic circuitry (e.g., MOS @ 77°K).

- For the microcapillary thermal interface technology, investigation of the feasibility of using interfacial liquids other than silicone oil, in particular liquid metals. Also, investigation of the use of plated reentrant microcapillaries to achieve **void-free soldering** of surfaces (perhaps using a gaseous flux). Finally, investigation of the utility of using the interface in a vacuum as a high-thermal conductance contact, i.e., a "vacuum chuck" which can operate in a high vacuum.

- Investigation of thermal spreading resistance from submicron transistors. In particular, at low temperatures the phonon mean free paths will become greater than the device size, causing locally ballistic heat transport which will greatly reduce thermal spreading resistance.

6.2.2. Electrical Considerations

Now that the thermal problems of high-performance ICs have been dealt with in this work, the question remains as to whether one can deal with the electrical problems of designing, for example, high-density bipolar VLSI circuits. Such ICs would consume hundreds of watts per cm^2. It is the author's opinion that this can be done. However, some problems need to be researched and solved. The key problem is the rather high current levels which would exist in the thin film conductors, particularly power busses. Electromigration is often cited as posing a limit to current levels, but in fact current densities of 2×10^5 A/cm^2 are readily achievable using modern aluminum alloys, and the problem is nonexistent for refractory metals at any reasonable temperature. Thin-film ohmic drops, rather than electromigration, will be more of

a problem in the design of high-power, large-area VLSI chips. The following are some initial research recommendations.

- Development of multilayer thin-film metallization techniques for integrated circuits, which have thick ground and power plane layers as well as sufficiently conductive signal transmission wires.

- Design of circuit techniques which are, to some extent, tolerant of ohmic drops and signal attenuation in long power and signal wires.

- Development of on-chip voltage regulators to avoid the problem of inductive noise on signal pins from the switching of nearby high-power chips. Note that this might not be necessary if a balanced-termination logic family such as ECL is used.

- Investigation of the use of cryogenic logic to take advantage of the lowered resistance of the metal lines.

6.3. Other applications

In addition to cooling integrated circuits, the following other applications of silicon heat-transfer microstructures are suggested:

- Rapid thermal cycling of thin films for materials testing. The micro-heat sinks have thermal time constants of a few milliseconds, hence millions of thermal cycles can be performed daily. If a test thin film were coated on top of a thin-film heater resistor which in turn sits on a silicon heat sink, the test film's response to multiple thermal stress cycles could be rapidly determined.

- Measurement of the thermal conductivity of thin films. Normally such measurements are difficult because the heat flux is limited by heat sink capability. The silicon micro-heat sink technology allows us to maintain spatially uniform heat fluxes of >1 kW/cm^2 through a test thin film (located between the heater resistor and the silicon heat sink substrate) without excessive temperature rise. The thermal conductivity of the test film could be most accurately deduced by incorporating a step in the test film.

References

1. Francois D'Heurle, "Electromigration and Failure in Electronics: An Introduction," *Proceedings of the IEEE,* Vol. 59, No. 10, October 1971, pp. 1409-1418.

2. J. D. McBrayer, R. M. Swanson, T. W. Sigmon, and J. Bravman, "Observation of Rapid Field-Aided Diffusion of Silver in Metal-Oxide-Semiconductor Structures," *Applied Physics Letters,* Vol. 43, No. 7, 1 October 1983, pp. 653-654.

3. George A. Sai-Halasz, Matthew R. Wordeman, and Robert H. Dennard, "Alpha-Particle-Induced Soft Error Rate in VLSI Circuits," *IEEE Transactions on Electron Devices,* Vol. ED-29, No. 4, April 1982, pp. 725-731.

4. Richard E. Matick, *Transmission Lines for Digital and Communications Networks,* McGraw-Hill, New York, 1969.

5. Richard E. Matick, *Computer Storage Systems and Technology,* John Wiley & Sons, New York, 1977, pp. 327, ISBN 0-471-57629-8.

6. E. E. Davidson, "Electrical Design of a High Speed Computer Package," *IBM Journal of Research and Development,* Vol. 26, No. 3, May 1982, pp. 349-361.

7. Krishna C. Saraswat and Farrokh Mohammadi, "Effect of Scaling of Interconnections on the Time Delay of VLSI Circuits," *IEEE Transactions on Electron Devices,* Vol. ED-29, No. 4, April 1982, pp. 645-650.

8. A. K. Sinha, J. A. Cooper, Jr., and H. J. Levinstein, "Speed Limitations due to Interconnect Time Constants in VLSI Integrated Circuits," *IEEE Electron Device Letters,* Vol. EDL-3, No. 4, April 1982, pp. 90-92.

9. Hadis Morkoc and Paul M. Solomon, "The HEMT: a superfast transistor," *IEEE Spectrum,* Vol. 21, No. 2, February 1984, pp. 28-35.

10. R. W. Keyes, "Physical Limits in Digital Electronics," *Proceedings of the IEEE,* Vol. 63, May 1975, pp. 740-767.

11. R. W. Keyes, "Limitations of Small Devices and Large Systems," in *VLSI Electronics: Microstructure Science,* Academic Press, 1981, ch. 5, ISBN 0-12-234101-5.

12. W. Anacker, "Josephson Computer Technology: An IBM Research Project," *IBM Journal of Research and Development,* Vol. 24, No. 2, March 1980, pp. 107-112.

13. R. C. Chu, U. P. Hwang, and R. E. Simons, "Conduction Cooling for an LSI Package: A One-Dimensional Approach," *IBM Journal of Research and Development,* Vol. 26, No. 1, January 1982, pp. 45-54.

14. S. Oktay and H. C. Kammerer, "A Conduction-Cooled Module for High-Performance LSI Devices," *IBM Journal of Research and Development,* Vol. 26, No. 1, January 1982, pp. 55-56.

15. E. A. Wilson, "True liquid cooling of computers," *Proc. 1977 National Computer Conference,* AFIPS Press, Montvale, NJ, 1977, pp. 341-348.

16. William M. Kays and Michael E. Crawford, *Convective Heat and Mass Transfer, 2nd ed.,* McGraw-Hill, New York, 1980, ISBN 0-07-033457-9. See in particular chapters 6 and 8 on laminar flow in tubes.

17. W. M. Kays and A. L. London, *Compact Heat Exchangers, 2nd ed.,* McGraw-Hill, New York, 1964.

18. A. H. Johnson, "Integrally Groove Semiconductor Chip and Heat Sink," *IBM Technical Disclosure Bulletin,* Vol. 14, No. 5, October 1971, pp. 1425.

19. T. H. Strudwick, "Silicon Tunnel Heat Sink," *IBM Technical Disclosure Bulletin,* Vol. 23, No. 2, July 1980, pp. 579-580.

20. A. L. London, "Heat Transfer Conditions in High Power Klystron Tubes," Tech. report 87-800-112, Varian Associates, October 1970.

21. D. B. Tuckerman and R. F. W. Pease, "High-Performance Heat Sinking for VLSI," *IEEE Electron Device Letters,* Vol. EDL-2, No. 5, May 1981, pp. 126-129.

22. D. B. Tuckerman and R. F. W. Pease, "Optimized convective cooling using micromachined structures," *Electrochemical Society Extended Abstracts,* 9-14 May 1982, pp. 197-198, Abstract # 125.

23. D. B. Tuckerman and R. F. W. Pease, "Ultrahigh thermal conductance microstructures for cooling integrated circuits," *32nd Electronics Components Conference Proceedings,* May 1982, pp. 145-149.

24. A. Amendola, C. A. Peck, and C. Prasad, "Cooling Structure for an Integrated Circuit Module," *IBM Technical Disclosure Bulletin,* Vol. 23, No. 2, July 1980, pp. 602.

25. J. R. Lynch, "Air and Liquid Drop Cooled Module," *IBM Technical Disclosure Bulletin,* Vol. 22, No. 1, June 1979, pp. 97-98.

26. L. D. Comerford, "Flip-Chip Bonding by Solder Filling of Capillaries," *IBM Technical Disclosure Bulletin,* Vol. 23, No. 5, October 1980, pp. 2146-2147.

27. R. C. Chu and U. P. Hwang, "Fluidized Thermal Interface," *IBM Technical Disclosure Bulletin,* Vol. 23, No. 2, July 1980, pp. 700-701.

28. D. B. Tuckerman and R. F. W. Pease, "Microcapillary Thermal Interface Technology for VLSI Packaging," *1983 Symposium on VLSI Technology,* 1983, pp. 60-61, Available from IEEE Service Center, Piscataway, NJ 08854.

29. Charles Kittel, *Introduction to Solid State Physics, 5th ed.,* John Wiley and Sons, New York, 1976, pp. 143-144.

30. C. Y. Ho, R. W. Powell and P. E. Liley, *J. Phys. Chem. Ref. Data,* , 1974, pp. I-588, Supplement 1.

31. D. G. Arasli and M. I. Aliev, "Influence of Defects and of the Interaction between them on Phonon Scattering in Heavily Doped Ge and Si Crystals," *Physica Status Solidi,* Vol. 21, 1967, pp. 643-649.

32. R. Gereth and K. Hubner, "Phonon Mean Free Path in Silicon Between 77 and 250°K," *Physical Review,* Vol. 134, No. 1A, 6 April 1964, pp. A235-A240.

33. Richard C. Joy and E. S. Schlig, "Thermal Properties of Very Fast Transistors," *IEEE Transactions on Electron Devices,* Vol. ED-17, No. 8, August 1970, pp. 586-594.

34. R. K. Shah and A. L. London, *Advances in Heat Transfer, Supplement 1: Laminar Flow Forced Convection in Ducts,* Academic Press, New York, 1978, ISBN 0-12-020051-1

35. W. M. Kays and A. L. London, *Compact Heat Exchangers, 2nd ed.,* McGraw-Hill, New York, 1964, pp. 14.

36. Robert W. Keyes, "Heat Transfer in Forced Convection Through Fins," Tech. report RC 9895, IBM Thomas J. Watson Research Center, March 1983, (Unpublished).

37. R. D. Cess and E. C. Shaffer, "Heat Transfer to Laminar Flow between Parallel Plates with a Prescribed Wall Flux," *Applied Scientific Research,* Vol. 8A, 1958, pp. 339-344.

38. A. P. Colburn, *Trans. Am. Inst. Chem. Engrs.*, Vol. 29, 1933, pp. 174-209.

39. A. Barba et al, "Numerical Study of Viscous Flow in Spirally Fluted Tubes," *Bull. Amer. Phys. Soc.,* Nov 1982, pp. 1180-1181, Abstract DA8.

40. G. A. Kemeny and J. A. Cyphers, "Heat Transfer and Pressure Drop in an Annular Gap with Surface Spoilers," *Journal of Heat Transfer,* May 1961, pp. 189-198.

41. A. L. London, private communication.

42. Kemeny and Cyphers, op cit, page 193.

43. A. L. London, "Coolant Path Hydraulic Design for High Power Twystron® Tubes," Tech. report, Varian Associates, October 1970.

44. Kays and London, op cit, Chapter 5.

45. R. G. Deissler, "Analytical Investigation of Fully Developed Laminar Flow in Tubes with Heat Transfer with Fluid Properties Variable along the Radius," Tech. report TN 4210, NACA, July 1951.

46. G. L. Pearson, W. T. Read, Jr., and W. L. Feldman, "Deformation and Fracture of Small Silicon Crystals," *Acta Metallurgica,* Vol. 5, April 1957, pp. 181-191.

47. R. K. Shah and A. L. London, "Effects of Nonuniform Passages on Compact Heat Exchanger Performance," *Journal of Engineering for Power,* Vol. 102, No. 3, July 1980, pp. 653-659.

48. Kenneth E. Bean, "Anisotropic Etching of Silicon," *IEEE Transactions on Electron Devices,* Vol. ED-25, No. 10, October 1978, pp. 1185-1193.

49. Don L. Kendall, "On Etching Very Narrow Grooves in Silicon," *Applied Physics Letters,* Vol. 26, No. 4, February 1975, pp. 195-198.

50. Don L. Kendall, "Vertical Etching of Silicon at Very High Aspect Ratios," *Ann. Rev. Mater. Sci.,* Vol. 9, 1979, pp. 373-403.

51. Product of Dynatex Corp., Redwood City, CA.

52. Boris W. Batterman, "Hillocks, Pits, and Etch Rate in Germanium Crystals," *Journal of Applied Physics,* Vol. 28, No. 11, November 1957, pp. 1236-1241.

53. R. J. Jaccodine, "Use of Modified Free Energy Theorems to Predict Equilibrium Growing and Etching Shapes," *Journal of Applied Physics,* Vol. 33, No. 8, August 1962, pp. 2643-2647.

54. Kurt E. Petersen, "Silicon as a Mechanical Material," *Proceedings of the IEEE,* Vol. 70, No. 5, May 1982, pp. 420-457.

55. Ernest Bassous, "Fabrication of Novel Three-Dimensional Microstructures by the Anisotropic Etching of (100) and (110) Silicon," *IEEE Transactions on Electron Devices,* Vol. ED-25, No. 10, October 1978, pp. 1178-1185.

56. A. G. Emslie, F. T. Bonner and L. G. Peck, "Flow of a Viscous Liquid on a Rotating Disk," *Journal of Applied Physics,* Vol. 29, No. 5, May 1958, pp. 858-862.

57. G. Wallis and D. I. Pomerantz, "Field Assisted Glass-Metal Sealing," *Journal of Applied Physics,* Vol. 40, 1969, pp. 3946-3849.

58. G. Wallis, "Direct-Current Polarization during Field-Assisted Glass-Metal Sealing," *J. Amer. Ceram. Soc.,* Vol. 53, 1970, pp. 563-567.

59. J. M. Brownlow, "Glass Related Effects in Field-Assisted Glass-Metal Bonding," Tech. report RC 7101 (# 30435), IBM, May 1978.

60. Y. S. Touloukain and C. Y. Ho (Editors), *Thermophysical Properties of Matter: the TPRC Data Series,* IFI/Plenum, New York, 1977, , Volume 13: Thermal Expansion of Nonmetallic Solids.

61. Richard S. Muller and Theodore I. Kamins, *Device Electronics for Integrated Circuits,* John Wiley and Sons, New York, 1977, pp. 32-33, ISBN 0-471-62364-4.

62. B. Glaser and Gerald E. Subak-Sharpe, *Integrated Circuit Engineering: Design, Fabrication and Applications,* Addison-Wesley, Reading, Mass., 1977, pp. 228.

63. F. Mohammadi and K. C. Saraswat, "Properties of Sputtered Tungsten Silicide for Integrated Circuit Applications," *Journal of the Electrochemical Society,* Vol. 127, No. 2, February 1980, pp. 450-454.

64. Howard H. Manko, *Solders and Soldering, 2nd ed.,* McGraw-Hill, 1979, ISBN 0-07-039897-6.

65. John Boucher and Fumio Wada, "The Resin-Bonded Diamond Blade for Dicing Hard, Brittle Materials," *Microelectronics Manufacturing and Testing,* Vol. 4, No. 2, February 1981, .

66. A. Murty Kanury, *Introduction to Combustion Phenomena,* Gordon and Breach Science Publishers, New York, 1977, pp. 63, Combustion Science and Technology Book Series Volume 2, ISBN 0-677-02690-0.

67. H. D. Baker, E. A. Ryder, and N. H. Baker, *Temperature Measurement in Engineering,* John Wiley and Sons, New York, 1963, pp. 71-72.

68. F. G. Hammitt, op cit, Chap. 2.

69. H. L. Oh, K. P. L. Oh, S. Vaidyanathan and I. Finnie, "On the Shaping of Brittle Solids by Erosion and Ultrasonic Cutting," *The Science of Ceramic Machining and Surface Finishing,* S. J. Schneider, Jr. and R. W. Rice, ed., National Bureau of Standards, Washington, D. C., May 1972, pp. 119-132, NBS special publication 348.

70. Jerry Lyman, "Packaging VLSI," *Electronics,* 29 December 1981, pp. 66-75.

71. Glaser and Subak-Sharpe, op cit, Chapter 10.

72. C. A. Neugebauer, "Soft Solder Fatigue in Power Semiconductor Packaging," Technical Information Series 82CRD151, General Electric Corporate Research and Development, June 1982.

73. W. T. Chen and C. W. Nelson, "Thermal Stress in Bonded Joints," *IBM Journal of Research and Development,* Vol. 23, No. 2, March 1979, pp. 179-188.

74. D. A. Spera and D. F. Mowbray, editor, *Thermal Fatigue of Materials and Components,* American Society for Testing and Materials, Philadelphia, 1976.

75. G. A. Lang, B. J. Fehder, and W. D. Williams, "Thermal Fatigue in Silicon Power Transistors," *IEEE Transactions on Electron Devices,* Vol. ED-17, No. 9, September 1970, pp. 787-792.

76. Nathan D. Zommer, Donald L. Feucht, and Richard W. Heckel, "Reliability and Thermal Impedance Studies in Soft Soldered Power Transistors," *IEEE Transactions on Electron Devices,* Vol. ED-23, No. 8, August 1976, pp. 843-850.

77. Donald R. Kitchen, "Physics of Die Attach Interfaces," *18th Annual Proceedings of Reliability Physics Conference,* IEEE, New York, April 1980, pp. 312-317.

78. Michael Lancaster (AMD Corp.), private communication.

79. A. W. Brunot and Florence F. Buckland, "Thermal Contact Resistance of Laminated and Machined Joints," *Transactions of the ASME,* April 1949, pp. 253-257.

80. N. D. Weills and E. A. Ryder, "Thermal Resistance Measurements of Joints Formed Between Stationary Metal Surfaces," *Transactions of the ASME,* April 1949, pp. 259-267.

81. A. M. Clausing and B. T. Chao, "Thermal Contact Resistance in a Vacuum Environment," *Journal of Heat Transfer,* May 1965, pp. 243-251.

82. W. A. Little, "The Transport of Heat between Dissimilar Solids at Low Temperatures," *Canadian Journal of Physics,* Vol. 37, 1959, pp. 334-349.

83. Wakefield Corporation product literature.

84. R. Defay and I. Prigogine, *Surface Tension and Adsorption,* Longmans, London, 1966.

85. Frank B. Kenrick, C. S. Gilbert, and K. L. Wismer, "The Superheating of Liquids," *Journal of Physical Chemistry,* Vol. 28, 1924, pp. 1297-1307.

86. Louis Bernath, "Theory of Bubble Formation in Liquids," *Industrial and Engineering Chemistry,* Vol. 44, No. 6, June 1952, pp. 1310-1313.

87. A. S. Tucker and C. S. Ward, "Critical State of Bubbles in Liquid-Gas Solutions," *Journal of Applied Physics,* Vol. 46, No. 11, November 1975, pp. 4801-4808.

88. R. J. Jaccodine and W. A. Schlegel, "Measurement of Strains at Si-SiO_2 Interface," *Journal of Applied Physics,* Vol. 37, No. 6, May 1966, pp. 2429-2434.

89. A. K. Sinha, H. J. Levinstein, and T. F. Smith, "Thermal Stresses and Cracking Resistance of Dielectric Films (SiN, Si_3N_4, and SiO_2) on Si Substrates," *Journal of Applied Physics,* Vol. 49, No. 4, April 1978, pp. 2423-2426.

90. D. Thebault and L. Jastrzebski, "Review of Factors Affecting Warpage of Silicon Wafers," *RCA Review,* Vol. 41, December 1980, pp. 592-611.

91. L. D. Yau, "Correlation Between Process-Induced In-Plane Distortion and Wafer Bowing in Silicon," *Applied Physics Letters,* Vol. 33, No. 8, 15 October 1978, pp. 756-758.

92. L. D. Yau, "Process-Induced Distortion in Silicon Wafers," *IEEE Transactions on Electron Devices,* Vol. ED-26, No. 9, September 1979, pp. 1299-1305.

93. David McFarland, Bert L. Smith, and Walter D. Bernhart, *Analysis of Plates,* Spartan Books, New York, 1972.

94. H. F. Wolf, *Silicon Semiconductor Data,* Pergamon Press, New York, 1969.

95. CRC Handbook of Chemistry and Physics, 54th ed., page F-22

96. W. A. Zisman, "Relation of the Equilibrium Contact Angle to Liquid and Solid Constitution," in *Contact Angle, Wettability, and Adhesion,* Frederick M. Fowkes, ed., American Chemical Society, Washington, D.C., 1964, ch. 1, Advances in Chemistry Series Volume 43.

97. Information about silicone diffusion pump fluids from Dow Corning Corp., Midland, Michigan., .

98. Robert L. Cottington, Charles M. Murphy, and Curtis R. Singleterry, "Effects of Polar-Nonpolar Additives on Oil Spreading on Solids, with Applications to Nonspreading Oils," in *Contact Angle, Wettability, and Adhesion,* Frederick M. Fowkes, ed., American Chemical Society, Washington, D.C., 1964, pp. 341-354, ch. 25, Advances in Chemistry Series Volume 43.

99. John F. O'Hanlon, *A User's Guide to Vacuum Technology,* John Wiley and Sons, New York, 1980, ISBN 0-471-01624-1

100. W. Crowder (IBM), private communication

101. R. Lefferts (Tektronix), private communication

102. Glaser and Subak-Sharpe, op cit, page 230

103. CRC Handbook of Chemistry and Physics, 54th ed., Page F-60.

104. Lester F. Spencer, "Electroless Nickel Plating - A Review," *Metal Finishing,* Vol. 72, No. 10, October 1974, pp. 35-45, Series of 4 articles (through Jan. 1975 issue

105. G. Salvago and P. L. Cavallotti, "Characteristics of the Chemical Reduction of Nickel Alloys with Hypophosphites," *Plating,* Vol. 59, July 1972, pp. 665-671.

106. Miles V. Sullivan and John H. Eigler, "Electroless Nickel Plating for Making Ohmic Contacts to Silicon," *Journal of the Electrochemical Society,* Vol. 104, No. 4, April 1957, pp. 226-230.

107. H. E. Austen and R. D. Fisher, "Internal Stresses of Electroless Metal Films on Single Crystal Silicon," *Journal of the Electrochemical Society,* Vol. 116, No. 2, February 1969, pp. 185-7.

108. L. W. Herron, J. Suierveld, and P. A. Totta, "Lowered Stress, Crack-Free Electroless Ni Deposits," *IBM Technical Disclosure Bulletin,* Vol. 23, No. 8, January 1981, pp. 3670.

109. CRC Handbook of Chemistry and Physics, 54th ed., Page B-94.

110. Frederick G. Hammitt, *Cavitation and Multiphase Flow Phenomena,* McGraw-Hill, 1980, ch. 3: Bubble Growth and Nucleation, ISBN 0-07-025907-0.

111. Barnes Engineering Co., Stamford, Connecticut