Thermal Performance Challenges from Silicon to Systems

Ram Viswanath, Technology and Manufacturing Group, Intel Corp.
Vijay Wakharkar, Technology and Manufacturing Group, Intel Corp.
Abhay Watwe, Technology and Manufacturing Group, Intel Corp.
Vassou Lebonheur, Technology and Manufacturing Group, Intel Corp.

Index words: power, thermal management, thermal design, packaging, heat sinks, thermal-interface materials

ABSTRACT

The demand for high-performance microprocessors has resulted in an escalation of power dissipation as well as heat flux at the silicon level. At the same time, the desire for smaller form-factor chassis and lower silicon operating temperatures is compounding the thermal challenge. Thermal design for a microprocessor can no longer be treated in isolation. Power and performance trade offs and smart circuit-design techniques are required to conserve power consumption. Materials and process improvements in packaging and heat-sink technology are required to minimize thermal resistance. The concurrent development and packaging of all these elements is critical to ensure that from a cost and availability perspective a viable thermal design solution space exists. This paper attempts to address this multidimensional problem, highlighting design and technology challenges encountered in mobile computers, desktops, and servers.

INTRODUCTION

The current trend in microprocessor architecture is to increase the level of integration (higher power), shrink processor size (smaller die), and increase clock speeds (higher frequency). Simply stated, this results in an increase in both the raw power as well as the power density on silicon. The drive to manage yield and reliability is resulting in the need for lower operating temperatures. This in turn translates to a shrink in temperature budgets for thermal design. Careful management of the thermal design space from the silicon to the system level is therefore critical to ensure a viable solution space for succeeding generations of processors.

The first section of this paper examines the trends in packaging and system-level thermal budgets. Design considerations of silicon such as power conservation features (clock gating, deep sleep, etc.) to optimize power dissipation and performance of the device are explored.

The second section deals with packaging technology developments and focuses on thermal interface materials and integration. The modulation of critical-material properties for optimizing thermo-mechanical performance is described.

The final section deals with thermal solutions at the system level. The discussion includes heat-sink technologies such as extruded, folded fin, and fan heat sinks. Limitations arising from chassis standards pertaining to airflow, board layout, heat-sink volumetrics, and acoustics are explored. Challenges with cost and scalability of new technologies such as vapor chamber heat sinks and high-power processors are highlighted. Key vectors relating to performance, cost, and form-factor trends in mobile, desktop, and server markets are examined.

MICROPROCESSOR POWER AND HEAT FLUX TRENDS

The insatiable demand for higher performance processors has lead to a steady escalation in power consumption across all the market segments, i.e., mobile and performance desktops as well as servers and workstations. Consider Figure 1 which shows the evolution of CPU power in the performance desktop market over the past decade.

It is seen that as the frequency scales higher over time, so does the power dissipation of the microprocessors. The improvements in process have been able to hold the power increase to reasonable levels, but it is definitely trended higher. A similar trend is reflected in the average heat flux (power dissipated per unit die area) on the processor, indicating a linear increase over time. This

is due to the fact that the power reduction obtained from architecture and process modifications is not commensurate with the scaling in die size, voltage, and frequency to support a cap in power consumption. In addition, the wider range of power and frequency offerings will enable performance and cost trade offs to be made between the various market segments.

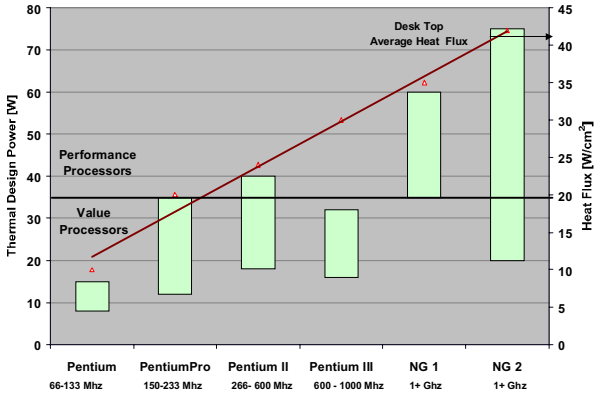
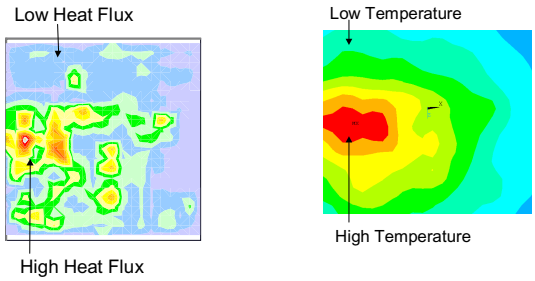


Figure 1: Power and heat-flux trends in the performance and value desktop processors

The need for higher performance and an increased level of functional integration as well as die size optimization on the microprocessor leads to preferential clustering of higher power units on the processor. In turn, this leads to a higher heat-flux concentration in certain areas of the die and lower heat fluxes in other regions on the die, which manifest themselves as large temperature gradients on the die (see Figure 2). This issue is becoming increasingly important as we deal with the emerging generation of microprocessor architectures. Simply stated, the thermal designs have to meet stringent heat-flux requirements that are significant multiples of the average heat flux at the silicon-package interface.



(a) (b)

Figure 2: Distribution of heat flux and temperature on the processor

A similar trend is seen in the mobile processor market segment. This market segment is typically constrained by battery life (~ 2-3 hours). The additional constraint is that the form-factor must be small and light to allow for portability. The desktop market is cost sensitive, and the mobile market is space and weight sensitive. These sensitivities place bounds on the effective power removal capabilities of the chassis. The long-term solution to the problem is addressed in a twofold manner:

- The design and architecture of the microprocessor must be such that it optimizes performance and power consumption.
- New cost-effective technologies in microprocessor and system packaging must be developed.

THE ISSUE OF SHRINKING TEMPERATURE BUDGETS

Why is Thermal Management a Critical Issue for Computing Systems?

There are two major reasons to maintain the operating temperature of a device at a certain level.

1. It is a well-known fact that the reliability of circuits (transistors) is exponentially dependent on the operating temperature of the junction. As such, small differences in operating temperature (order of 10–15°C) can result in a ~2X difference in the lifespan of the devices.
2. The other factor is the speed of the microprocessor. At lower operating temperatures, due to reduced gate delay, microprocessors can operate at higher speeds. A secondary effect of lower temperatures is related to a reduction in idle power dissipation (also known as leakage power) of the devices, which manifests itself as reduction in overall power dissipation. These two factors combined dictate the operating temperature of devices as a function of the speed of the device.

The next topic of discussion is thermal design and its associated complexities. In order to simplify this discussion, it is useful to introduce a metric known as thermal resistance, θ_{ja}, described as

$$T_j - T_a = \theta_{ja} * Power = R_{ja} * HeatFlux \qquad (1)$$

where T_j is the temperature of the device and T_a is the ambient temperature (in the vicinity of the device). The heat flux is the power dissipated per unit area (or volume) and is a metric that signifies the degree of localized power concentration, and R_{ja} is the thermal resistance

normalized over a unit area (or volume). Figure 3 shows the typical thermal resistance budgets for the emerging generation of processors. If the heat-removal mechanism is related to surface area (as in interfacial resistance), R_{x-y} is normalized over the wetted area. If the heat-removal mechanism is three-dimensional (as in convection), R_{x-y} is normalized over volumetrics.

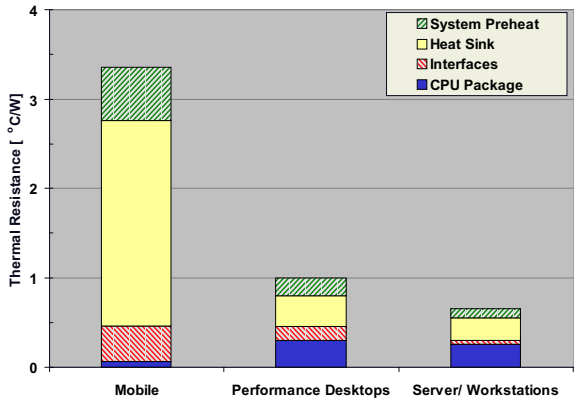


Figure 3: Typical thermal-resistance partitioning (across market segments)

From a purely technical standpoint, the lower the thermal resistance budgets, the more challenging the thermal design, and the higher the power capacity of the design. Additional metrics such as the volumetrics of a system and cost budgets add to the complexity and challenges of thermal design. Therefore the options available for improving a design lie in

- management of power consumption by the processor
- management of the design and technology elements to meet the individual thermal-resistance budgets

The next section focuses on power and performance-management techniques and tradeoffs.

DESIGN CONSIDERATIONS FOR POWER MANAGEMENT AND PERFORMANCE OPTIMIZATION

Power consumption trends on the microprocessor are becoming an increasing area of concern due to the complexities in packaging technologies as well as system thermal design and cost. As a result, there is increasing emphasis on microprocessor architecture and design to contain and manage processor power against performance as well as die area. In order to facilitate this discussion, it is important to understand the nomenclature used when defining power consumption on the microprocessor. This is shown in Table 1and schematically illustrated in Figure 4.

Table 1: Nomenclature and usage of power specifications

Parameter	Design Usage	Description
Maximum Power, P_{max}	Power Delivery Design	Maximum power drawn under normal operating conditions, worst case (Vcc,T) corner, executing worst case (synthetic) instruction set. Time duration of sampling is << O(thermal time constant).
Thermal Design Power, P_{tdp}	Thermal Design	Maximum sustained power, across a set of realistic applications, drawn under normal operating conditions, nominal Vcc and realistic ambient (use) temperature. Time duration of sampled data set is O(thermal time constant).
Active Power, P_{active}	Mobile Battery Life	Thermal design power time averaged over a period of time >> O(thermal time constant).
Idle Power	Thermal and Power Delivery Design /Battery Life	Power consumed in quiescent states where there is little or no clock activity. Examples are the sleep states of the processor such as deep/deeper sleep, stop clock, AutoHalt and so on. Stop clock power (STOPCLK# asserted, CLK not toggling) measured in nominal and worst case corners.

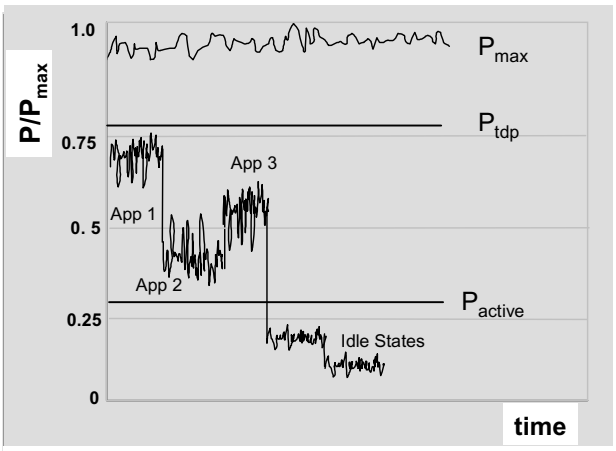


Figure 4: Example of microprocessor power consumption profiles

Thermal Design Targets: Power delivery designs are typically based on the theoretical maximum power that is drawn by the processor, which is based on a synthetic (power virus) code. The theoretical maximum power drawn is based on a synthetic code that is designed to use resident data from the on-chip caches (L1 or L2). The pipelines and queues are maintained full to the best possible extent. Given the superscalar and superpipelined architectures of the microprocessor, such activity could conceivably occur over brief bursts of time, but would not likely be sustainable over long periods. Furthermore, if the thermal designs are done to a lower power target (P_{tdp} for example) than the maximum power, the thermal capacity of the system may be able to support temporary bursts of power consumption over short durations, without violating the CPU thermal specifications.

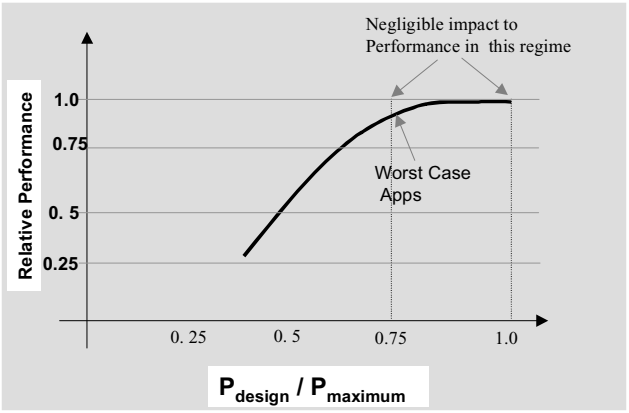


Figure 5: Illustration of impact of thermal design power on processor performance

Figure 5 shows a schematic illustration of the impact on system performance as a result of the thermal design power being lower than the maximum power. It is necessary to collect empirical data of this nature to quantify the impact on processor performance and determine the threshold power that should be used for thermal design. For example, thermal designs targeting 75% of the maximum power may have little or no impact on system performance.

Process Scaling: The most obvious power reduction is achieved through process technology optimization. Since $P \sim CV^2f$, significant reductions can be achieved at a given frequency by operating at a lower voltage as well as through a reduction in total capacitance (by reducing the die size).

Circuit Design: There continues to be a significant emphasis to manage power through design techniques. Design techniques such as clock gating and functional unit block power downs are used to conserve active power consumption on maximum power applications. The key is to ensure that the power savings are achieved without adversely impacting the die size or performance of the processor. Speculative execution techniques can provide a significant performance advantage but at the cost of increased power consumption. Therefore, power/ performance trade-off studies are carried out to ensure that the end result is a net gain.

Smart Voltage Regulation: The output voltage of the voltage regulator is a function of the load on the die. When the current drawn from the power supply is high, the output voltage droops. At zero load (current), the voltage delivered to the processor is highest. Classical designs involved maintaining Vcc flat over the entire range of current load. New voltage regulator designs are able to respond with dynamic voltage outputs that guarantee safe microprocessor operation with power savings. For example, the use of this method has shown to provide a 10-12% reduction in CPU thermal design power in the mobile environment. In addition, the use of advanced VR designs can result in idle power savings as well as improved battery life.

The following section deals with packaging-technology development with a focus on thermal interface materials.

PACKAGING-LEVEL SOLUTIONS

The packaging technology for microprocessors has primarily moved towards flip-chip attach for interconnecting the active side of silicon to an organic substrate. The substrate can be socketed in the case of Pin Grid Array (PGA) and surface mounted in the case of Ball Grid Array (BGA) packages. As is typical in flip-chip packaging, the primary mode of heat removal is from the back surface of the silicon. The thermal energy is removed ultimately by a heat sink to the surrounding ambient.

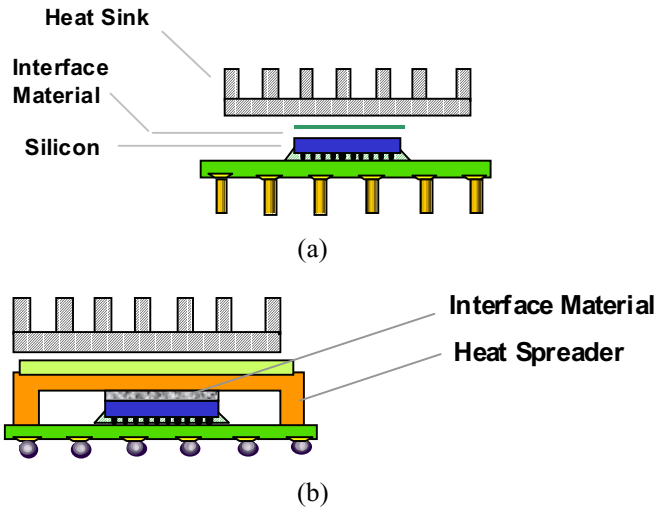


Figure 6: Schematics of FC-xGA1 and FC-xGA2 packaging architectures

The choice of packaging architecture used depends on the total power of the processor as well as the heat-flux density. Two basic architectures are identified: a) FC-xGA1 dealing with low-power processors, and b) FC-xGA2 dealing with medium- to high-power processors. The term xGA could stand for either PGA or BGA, and refers to the next level of interconnect. Figure 6 shows the basic implementation of these two architectures.

FC-xGA1 deals with directly interfacing the die to the heat sink and therefore involves the design and development of thermal-interface materials such as grease and phase-change films and associated retention mechanisms. FC-xGA2 architecture is designed to be scalable and meet the demands of medium- to high-performance (power) processors. These involve the integration of a heat spreader to the back surface of silicon using thermally conductive gels or epoxies. Thermal-interface materials are typically made up of a polymer matrix in combination with highly thermally conductive fillers (metal or ceramic). The materials technology is broadly classified into categories such as Thermal Epoxies, Phase Change Materials (PCM), Thermal Greases, and Gels.

GENERAL CONSIDERATIONS IN THE CHOICE OF MATERIALS FOR PACKAGING

The singular metric used to classify and select interface materials is the thermal resistance across two mated surfaces and is described by

$$\theta_{j-x} = \frac{T_j - T_x}{Power} \qquad (2)$$

Thermal test apparatus are available to characterize the performance of materials in a standalone fashion [1]. These methods are typically used to screen and rank order multiple materials based on thermal performance. The actual value obtained from screening setups may be different from in-situ performance of the same materials in a package due to sensitivity to surface finish, interface pressure, and so on. Nevertheless, this is an invaluable quick turn tool during the material selection process. A typical configuration of this tool is shown in Figure 7.

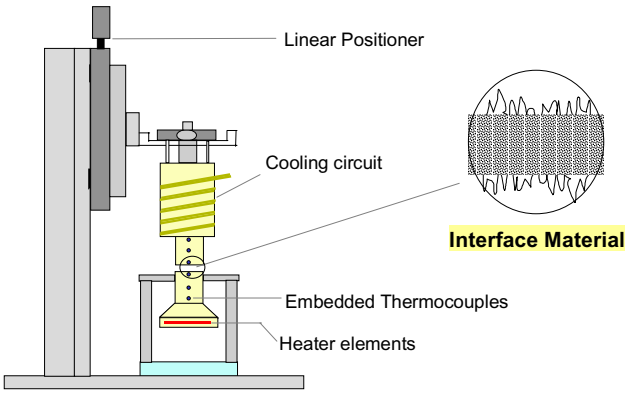


Figure 7: Schematic of set-up used to screen interface materials

In the micro-scale, the interface resistance can be expanded into the following entities:

$$\theta_{j-x} = \theta_0 + \frac{L}{k_m A_w} \qquad (3)$$

This equation is a simplified formulation of Fourier's law of diffusion [2], wherein θ_o is the contact resistance between the material and the mated surfaces, L is the thickness of the interface, k_m is the bulk thermal conductivity of the material, and A_w is the wetted surface area. Figure 8 shows a plot of this dependency.

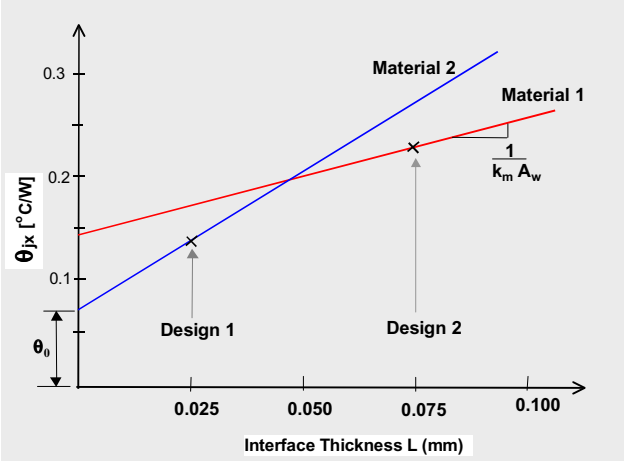


Figure 8: Schematic plot of dependence of thermal resistance to thickness and bulk material conductivity

At small interface thicknesses, it is seen that the choice of the interface material is a function of both the contact resistance and the bulk conductivity. The particular material chosen depends on the relative magnitude of these two entities at the target design thickness. It can also be inferred that for cases of low-interface thickness (< 0.1mm), the thermal contact resistance at the mating surfaces is a dominant factor [$\theta_o \cong L/(k_m A_w)$]. Figure 9 shows the actual thermal conductivity derived from measurements of an interface material over a range of thicknesses. It can be inferred that for cases of high-interface thickness (large interfacial thicknesses), $\theta_{j\text{-}x}$ approaches the value predicted by the bulk material conductivity [$\theta o << L/(k_m A_w)$].

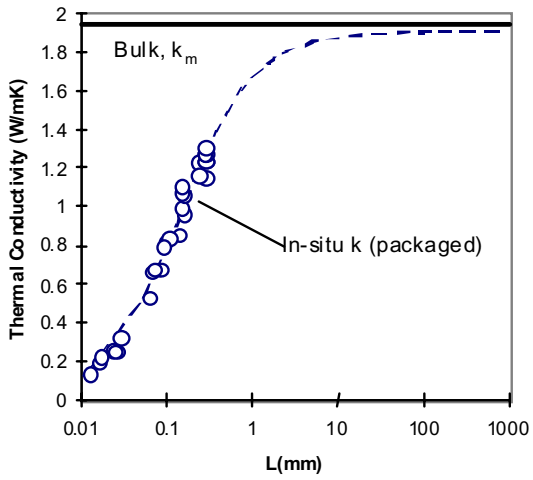


Figure 9: Variation of in-situ thermal conductivity of material (in a package) with thickness

Performance considerations as well as cost and manufacturability concerns, inevitably results in tradeoffs that are made during interface material selection. Some of these attributes are enumerated in the following section.

Choice of Matrix: Matrix selection is typically driven by its compatibility with filler, its mechanical properties, its ability to wet the mating surfaces, and its viscosity. The maximum filler loading that can be achieved is dictated by the thermodynamic wetability of the filler by the matrix and by the polymer viscosity. The polymer matrix also allows tailoring of the desired mechanical properties of the interface material under use conditions. Epoxy resins are used when high modulus and adhesion are targeted, whereas silicones are used for low-modulus and stress-absorbing applications. Lower surface energy materials are used to act as a matrix since they improve the wetability at the mating surfaces. Common examples of the use of silicones are the polymer materials.

Choice of Filler: The key ingredient in the interface material is the filler, which is responsible for heat conduction. The fillers are dispersed in a polymeric matrix, which typically has poor thermal conductivity, for handling and processability. The important filler properties are bulk thermal properties, morphology (size and shape), and distribution. To reduce the contact resistance, filler surface treatments (coating) are also critical for ensuring optimum filler and matrix thermodynamic wetting. Ceramic powders such as alumina and magnesium oxide are commonly used due to their lower cost and dielectric properties. Further thermal enhancement is achieved through more conductive particles such as aluminum nitride or boron nitride. These fillers provide a five to tenfold improvement in bulk thermal conductivity, but due to more elaborate manufacturing techniques, cost ten to one hundred times that of their ceramic counterparts. For higher performance, metal particles such as silver and aluminum are used. Silver is chosen mainly for its very high bulk thermal conductivity. Aluminum provides a balance between the bulk thermal conductivity and density (high-volume loading can be achieved because of low specific gravity).

Design Considerations: Issues such as physical design tolerances, positive pressure at the interface, warpage, and tilt and flatness of the mated surfaces have a direct influence on the thickness of the interface as well as the degree of wetting. Of particular interest is the tradeoff between the flatness (macroscopic) of the surface and the cost of machining. Warpage issues can be alleviated by the choice of assembly materials to lower the processing temperature as well as the coefficient of thermal expansion (CTE) mismatch between materials. Interface pressure becomes a key factor with collapsible materials; this is controlled through design.

Surface Finish: The interaction of filler particles with micro-structural asperities at the mated surfaces determines the degree of compaction and wetting at the interface.

Manufacturability Considerations: Materials in a semi-solid or liquid state need to be dispensable. A typical tradeoff is that a higher degree of filler loading (to reduce thermal resistance) translates into increased viscosity of the material, which in turn affects the manufacturing throughput. The degree of voiding at the interface has a significant impact on thermal performance, since voids are air gaps that act as conduction heat transfer barriers.

Reliability Considerations: Microprocessor packages are designed to survive field-use conditions typically over seven to ten years. Packages are subject to a range of stress suites to ensure that the device meets performance specifications over the lifespan of operation. An example of a stress suite is shown in Table 2.

Table 2: An example of environmental stresses imposed on a microprocessor package

Stress Parameter	Example Condition
Power Cycle	7500 cycles, 25 °C to 95 °C
Temperature Cycle	-55 °C to 125 °C
High Temperature Bake	125 °C for 168 hours
Mechanical Shock	50G sinusoidal wave with 250gm heat sink clipped
Temperature/Humidity Soak	55 °C/ 85% RH

FC-XGA1 PACKAGING ARCHITECTURE

This architecture encompasses packaging solutions for low- and medium-power microprocessors, predominantly encountered in the value and mobile processor market segments. As indicated earlier, the technology involves interfacing the heat sink to the die through a compliant interface material. The next section discusses in detail the technical elements of three classes of materials that are used in this architecture.

Elastomeric Thermal Pads

This class of materials (also known as gap-filler pads) is used to improve heat dissipation across large gaps, by establishing a conductive heat-transfer path between the mating surfaces. Thermal pads are typically 200 um to 1000 um thick and are popular for cooling low-power devices, such as chipsets and mobile processors. The pad consists of a filled elastomer, with filler materials ranging from ceramic to boron nitride for varying thermal performance. Metal particles are seldom used due to the risk of dislodged particles resulting in electrical shorts. Another key requirement is that the pads need to be compliant: They should be capable of being compressed to within 25% of their total thickness. This is necessary due to the tolerance variation of large gap situations. The compressibility ensures that the pads can absorb the tolerance variation in assemblies. The tradeoff therefore is that the increase in filler materials (for lower thermal resistance) results in hardening of the pad and hence reduced compliance. Typical failure mechanisms are increased thermal resistance due to inadequate pressure or loss of contact at one or more surfaces. The thermal performance is also sensitive to the contact pressure at the mated surfaces. Some thermal pads have a thin layer of pressure-sensitive adhesive (PSA) applied to promote adhesion at the interfaces. Nevertheless, the constraints discussed above significantly limit the thermal performance achievable with this class of materials.

Thermal Greases

Thermal greases offer several advantages over pads, including the ability to conform to the interfaces. They require no post-dispense processing (e.g., no cure) and they have higher effective thermal conductivity compared to other classes of materials. Greases have been used very successfully in combination with various packaging form-factors and have shown excellent performance. However, certain design and environmental considerations can preclude the use of thermal greases. The schematic set-up of package and heat sink used in the evaluations is shown in Figure 10.

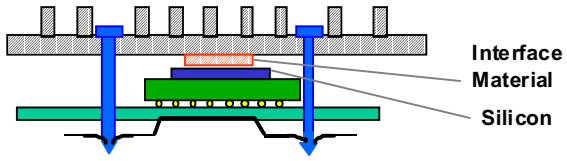
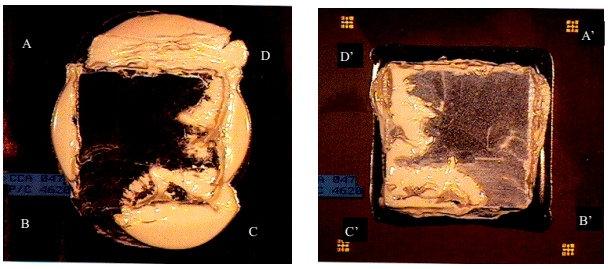


Figure 10: Schematic of the set-up used in thermal grease and phase-change film evaluations

This configuration was retained in a retention module mounted on a base board. The entire assembly was subjected to a range of reliability stresses. Thermo-mechanical jeopardy to the processor was identified under certain environmental stresses. The stress conditions and failure mechanisms are examined below.

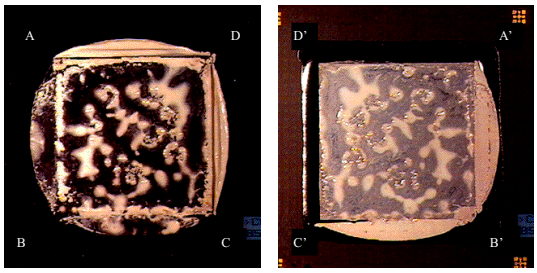
Power Cycling: This is a loss of material due to a phenomenon called "pump-out" (shown in Figure 11). Under cyclic loading, extensive thermo-mechanical stresses exerted at the interface because of the relative motion (flexure) between the die and the base of a heat-sink lead to loss of grease material from the interface.



Heat Sink Surface Die Back Surface

Figure 11: Typical illustration of thermal grease pump-out

Thermal Bake: Under high-temperature bake, the formulation chemistries utilized in typical thermal greases result in separation of the polymer and filler matrix due to the migration of the polymer component. The separation and loss of polymeric material could result in poor wetability at the interfaces, resulting in a increase in thermal resistance, also known as "dry-out" (shown in Figure 12). This phenomenon is strongly dependent on the temperature of the material with higher temperatures resulting in accelerated degradation.



Heat Sink Surface Die Back Surface

Figure 12: Typical illustration of thermal grease phase separation and dry out

The failure mechanisms encountered are a strong function of the thermal grease operating temperature and the number of on/off cycles that the processor assembly has been subjected to. The rate of thermal degradation is also dependent on the surface finish of the mating surfaces (heat spreader surface vs. back side of silicon). The pump-out mechanism and phase separation mechanisms have an exponential dependence on temperature, with a twofold increase in degradation for every 10 °C increase in average operating temperature of the interface material. Data collected also indicate that for power cycling, the assembly between 0 and 100 °C over 7500 cycles results in a four to sixfold increase in thermal resistance compared to a negligible increase in resistance for a 0 to 80 °C exposure over 2500 power cycles (see Figure 13).

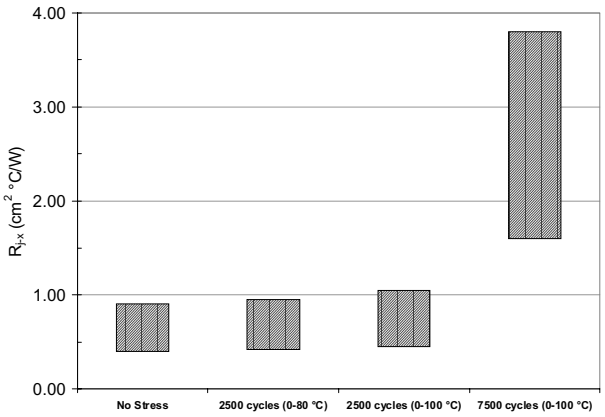


Figure 13: Impact of environmental stress conditions on thermal grease performance

Mechanical Shock and Vibration: The second area of concern is related to retention of a heavy heat-sink mass interfaced to the die through a compliant material such as grease. Data collected on mechanical shock and vibration of heat-sink masses between 200-250 grams indicate that the retention of the heat sink to the processor is critical to prevent die damage. During shock testing, relative motion between the heat sink and die could lead to mechanical damage to the die surface, with the corners of the die being highly susceptible to damage.

This fail mechanism is strongly dependent on the design of the heat-sink retention feature as well as the mass of the heat sink used. Typically the heat-sink mass (or volume) is proportional to the power dissipation of the processor; a heavier heat sink is required to cool higher processor power. Data collected indicate that in the case of lighter die loading, die damage due to mechanical shock is not a concern.

In summary, thermal grease-based materials are recommended for applications at lower operational temperatures (to alleviate phase separation), lower die loading (to alleviate mechanical damage), and lower power cycling requirements (to alleviate pump-out). However, the limitations in the use of thermal grease triggered the development of an alternate material described in the following section.

Phase-Change Films

Phase-change films (PCFs) are a class of materials that undergoes a transition from a solid to semi-solid phase with the application of heat; The material is in a liquid phase under die-operating conditions. This class of materials offers several advantages including the ability to conform to profiles of the mating surfaces, no post-dispense processing (e.g., no cure), and ease of handling and processing due to its availability in a film format.

However, from a formulation perspective, the polymers and filler combinations that can be utilized impose limitations on the thermal performance of these materials.

PCFs are typically a polymer/carrier filled with a thermally conductive filler, which change from a solid to a high-viscosity liquid (or semi-solid) state at a certain transition temperature. The choice of materials is tailored such that the transition occurs below the operating temperature of the die. Key advantages of PCFs are related to their ability to conform to surfaces and their wetting properties, which significantly reduces the contact resistance at the different interfaces. These materials usually are reinforced with a fiberglass mesh, which acts as a core, providing mechanical rigidity. Due to this composite structure, PCF materials are able to withstand mechanical forces during shock and vibration, protecting the die from mechanical damage. The semi-solid state of these materials at elevated temperatures resolves issues related to "pump-out" under thermo-mechanical flexure. Typically, dispense processes required for thermal greases are throughput limiters. The manufacturing throughput of the assembly line is greatly improved since PCF's can be preattached to the base of a heat sink or heat spreader using a pick-and-place operation.

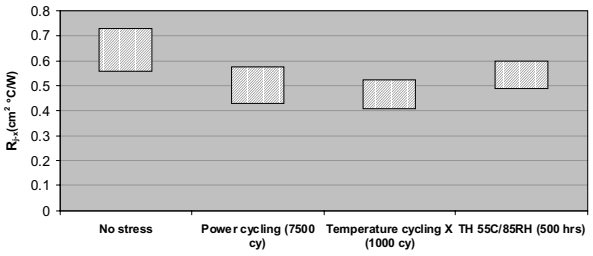


Figure 14: Performance of a PCF material through reliability stresses

The reliability of the PCF material assembled on the processor package was demonstrated through various stress suites. Thermal performance at time zero and post-reliability stress testing is shown in Figure 14. On average, it was observed that the thermal performance improves over the course of reliability stressing. Since the material is in a softened state during these stresses, the compressive loading at this interface resulted in a decrease in thickness as well as improved wetability (conformance) with the surface irregularities.

The following section describes the salient features of the FC-xGA2 packaging architecture.

FC-XGA2 PACKAGING ARCHITECTURE

The FC-xGA2 architecture is designed to be scalable to meet the demands of medium- to high-performance (power) processors. The integration of the heat spreader on the die required the development and certification of a thermally conductive polymer, a heat spreader, and an adhesive sealant.

The heat spreader helps in improving the diffusion of heat flux from the smaller die area to a much larger surface area. This in turn translates to improved thermal performance of the heat sink. Figure 15 shows a plot of the reduction in total thermal resistance due to improved spreading from three different heat spreader materials. Since copper has a higher thermal conductivity than AlSiC, it provides roughly 0.1°C/W lower thermal resistance due to improved heat spreading. The last curve shows a heat spreader with a hypothetical thermal conductivity of 2000 W/mK. The reduction of qja from interface material improvement is asymptotic. It is therefore advantageous to use high-conductivity heat spreaders after the interface resistance has been optimized. The desired design trend is suggested by the grey arrows shown in Figure 15.

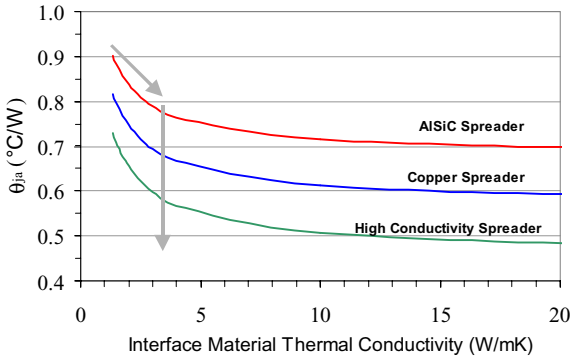


Figure 15: Dependency of thermal resistance with heat-spreader material and thickness

The architecture precluded the use of thermal greases due to the thermo-mechanical failure mechanisms. Phase-change materials were precluded since they did not meet the stringent thermal-performance requirements. In addition, the need for a positive compressive load at the interface imposes limitations on the design of this packaging architecture. The highly conductive and commonly utilized metal-filled epoxy thermal polymers could not be used here because of several major obstacles. The high-modulus nature of these materials leads to delamination at the interface due to thermo-mechanical stresses. In addition, localized phase

separation within the material resulted in high-contact resistance.

In order to overcome these technological issues, novel chemistries were aggressively investigated. As a result, a thermally conductive, low-modulus gel was developed. The gel is typically a metal particle (Al) or ceramic particle-filled (aluminum oxide, zinc oxide, etc.) silicone polymer with low cross-link density. It combines the properties of both a grease and a cross-linked polymer; i.e., it is a grease that can be slightly cured. Before cure, this material has properties similar to a grease: It has high-bulk thermal conductivities (2-5 W/mK) and conforms well to surface irregularities upon dispense and assembly. Post-cure, this material becomes a lightly cross-linked polymer with significantly lower modulus than epoxy systems. The cross linking reaction provides a high enough cohesive strength to the gel in order to circumvent the "pump-out" issues during temperature and power cycling. The modulus is maintained low enough (MPa range compared to GPa range of epoxies) so that the material can still absorb thermo-mechanical stresses to prevent interfacial delamination. The low-surface energies characteristic of silicones enable good wetting of the mated interfaces, which contribute to minimization of thermal resistance.

Figure 16 shows the evolution of thermal-interface materials to satisfy the continuously shrinking thermal budgets. The application of greases and elastomeric pads is restricted to low- and medium-power devices due to their inherently high-thermal resistance and limitations arising from reliability concerns. Several new material technologies including high-performance PCFs and metal particle-filled gels have been developed and integrated into the FC-xGA2 architecture to deal with high-power devices. Continued development in this area is necessary to satisfy the insatiable performance demands of the next generation of processors.

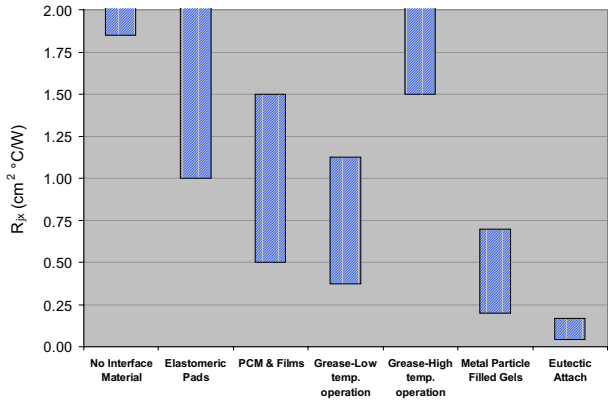


Figure 16: Thermal resistance of typical thermal-interface materials

SYSTEM-LEVEL SOLUTIONS

The primary goal of the system-level thermal solution is to extract the heat from the processor package and discharge it to the ambient air external to the chassis. Air cooling is the most widespread means of system-level cooling in desktop, workstation, and server segments. One or more system fans is employed to move the air within the chassis. Some of the parameters that affect the system thermal design are fan flow rate, acoustic limitations, ambient temperature, and the heat-sink volume.

General Considerations in System Thermal Design

Air cooling employs convection of heat from the heat sink to the ambient air. Figure 17 shows a typical rectangular fin heat sink.

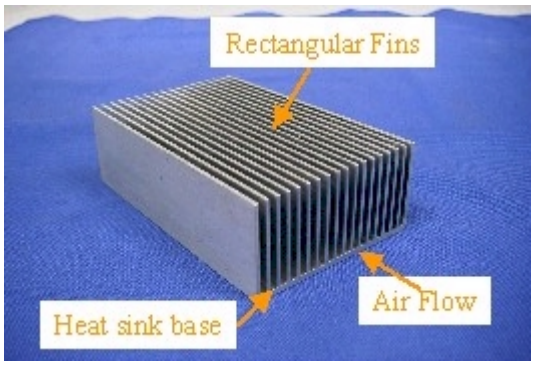


Figure 17: Rectangular fin heat sink

A good discussion on heat-sink designs can be found in [3]. The amount of heat that can be transferred by convection from the heat sink can be estimated using Newton's law of cooling:

$$Power = hA(T_s - T_a) \qquad (4)$$

where h is the heat transfer coefficient, T_s is the temperature on the surface of the heat sink, T_a is the ambient temperature, and A is the total surface area of the heat sink. Equation (4) can be rearranged to define the heat-sink thermal resistance as follows:

$$\theta_{sa} = \frac{T_s - T_a}{Power} = \frac{1}{hA} \qquad (5)$$

It is apparent from these equations that in order to increase the total heat transfer from the heat sink we must increase one or more of the parameters h, A, or T_s-T_a.

Each of these factors is briefly discussed in the following sections.

Surface Area

An increase in the heat-sink surface area increases the amount of heat that the heat sink can release to the ambient air stream. The heat-sink surface area can be increased either by increasing the number of fins or by modifying the shape of the fins (i.e., using dimpled or wavy fins instead of rectangular fins) [4]. However, increasing the heat-sink surface area results in an increase in the pressure drop across the heat sink. This is because the viscous shear stress acts over a larger area creating a larger frictional force. The system fan must be capable of generating a large enough pressure head to overcome the frictional resistance to the flow of air across the heat sink.

Heat Spreading

The temperature difference between the heat-sink surface and ambient air depends on the efficiency of heat spreading in the heat-sink base and fins. The amount of heat spreading depends primarily on the thermal conductivity of the heat-sink material and on the heat-sink geometry. For example, copper heat sinks have better spreading than aluminum heat sinks. Similarly, spreading is much better in a heat sink with a thicker base and thicker fins. Enhancing the heat spreading in the heat- sink base through the use of a vapor chamber adds to the total cost of the system cooling solution.

Air Flow and Pressure Drop

The heat transfer coefficient on the heat-sink fins depends primarily on the air flow rate, the spacing between the fins, and the flow regime (i.e., laminar or turbulent) that exists on the heat-sink fins. A higher air flow results in higher heat transfer coefficients and a correspondingly higher pressure drop. The heat-transfer coefficient also depends on whether the flow is fully developed.

Fin spacing and the air velocity can be used to determine if the flow is fully developed. In a fully developed flow, the boundary layers growing on adjacent fins merge within the heat sink. Closely spaced fins and lower air flow rates, which cause thicker boundary layers, will result in fully developed flow and lower heat-transfer coefficients. One way to break the growth of the boundary layer, thereby preventing the onset of a fully developed flow, is to use pin fins. Pin fins are usually formed by machining cross cuts across rectangular fins as shown in Figure 18. The gap between the pin fins serves to break the boundary layer; a new boundary layer develops on each downstream fin resulting in a higher heat-transfer coefficient. This is usually accompanied by an increase in the pressure drop across the heat sink. The flow in and around the heat sink showing the development of the velocity boundary layer as well as the recirculation zone in the wake of the heat sink can be seen in Figure 19. The temperature distribution in the heat sink as well as the thermal boundary layer effects are seen in Figure 20.

Increasing the fin thickness would lead to a reduction in the number of fins and heat-transfer area. This implies that there would be an optimum fin thickness at which the increased heat spreading would offset the contribution from the decreased fin area to provide the maximum heat transfer from the heat sink. Any increase in finned surface area or change in heat-sink base material from aluminum to copper results in an increase in the heat-sink weight.

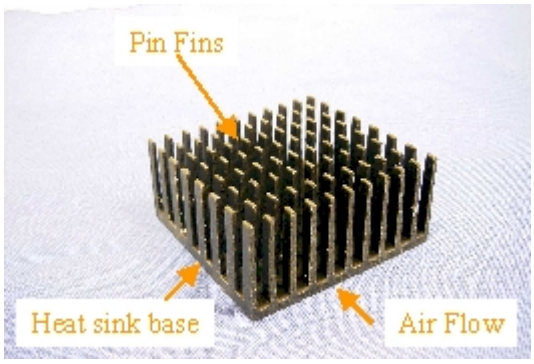


Figure 18: Pin fin heat sink

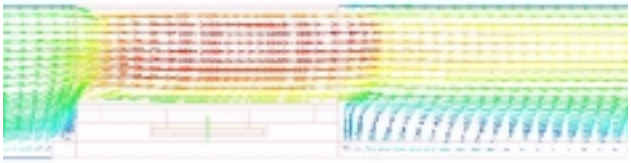


Figure 19: Velocity vectors around a Intel® Pentium® processor heat sink

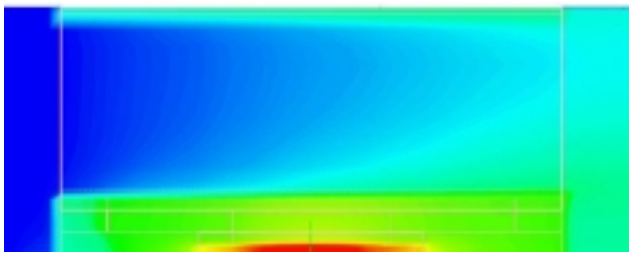


Figure 20: Temperature contours on a Intel® Pentium® processor heat sink

Acoustics

As discussed previously, any enhancement to the heat transfer from the heat sink is usually accompanied by an increase in the pressure loss across the heat sink. In order to overcome the pressure loss, a larger fan may be required. Increasing the air flow rate to increase the heat-transfer coefficient may also require a larger fan. Increased flow rates and larger fans typically result in increased fan noise. Noise attenuation schemes and larger system fans add to the total cost of the cooling solution.

Management of Ambient Temperature

In addition to selecting an optimum heat sink for cooling the processor, the system thermal designer must also consider the layout of the motherboard. In most desktop systems, one or two fans are used to provide air cooling for the processor and other auxiliary components such as memory chips, chipsets, graphics cards, etc. Consequently, the air gets preheated by these components before it reaches the processor heat sink. One simple technique to eliminate or reduce preheating effects is to supply outside ambient air to the processor by using a duct. Typically with this implementation, a second fan is required to provide air flow to cool the other components such as the chipset, memory, and graphics devices on the motherboard. The addition of a duct and a dedicated fan will increase the complexity and cost of the cooling solution.

Motherboard Layout for Optimum Thermal Design

With the advent of multiprocessor systems, there is a greater need to control the ambient temperature local to the processor. In a dual-processor system, it is sometimes necessary (due to electrical layout requirements) to place the second processor downstream of the primary processor. The second processor is therefore in the shadow of the first processor, and the ambient temperature local to the downstream processor may be as much as 10 to 15 °C higher than the upstream processor. One method to alleviate this problem is to lay out the processors in staggered rather than inline fashion on the motherboard.

Thermal Designs in Compact Chassis

The system thermal design for the mobile market segment (i.e., laptops and notebooks) poses a special set of challenges. There are severe space and weight constraints on the design of mobile products. Traditionally, mobile processors have been cooled via natural convection and radiation. With the increase in power dissipation, fans have been added to the notebook chassis to circulate the air and enhance the component cooling. In addition, heat pipes or heat spreaders have been used to transport the heat away from the processor to dissipate it at a remote location. One technique uses a copper plate to spread the heat over a large plate mounted just under the keyboard [5,6]. Natural convection from the keyboard is used to dissipate the heat to the ambient air. A second technique utilizes a heat pipe to transport the heat from the processor through the hinges to the back surface of the display panel. A third concept utilizes a remote heat exchanger with a dedicated fan. A heat pipe is used to transport the heat from the processor module to the remote heat exchanger, which is typically located near the outer wall of the laptop chassis. Figure 21 shows an example of this implementation.

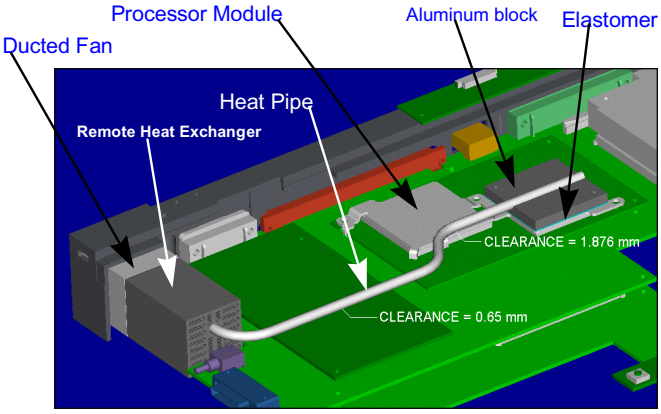


Figure 21: Remote heat exchanger concept for mobile Intel® Pentium® II processor module

HEAT-SINK TECHNOLOGY: PERFORMANCE AND COST

Active Versus Passive Heat Sinks

Heat sinks are usually classified as active or passive. Active heat sinks (see Figure 22) consist of a heat sink with a fan mounted directly to the heat sink. In an active heat sink, the fan blows air on the fins and base of the heat sink and provides cooling via air impingement. The use of active heat sinks is widespread in desktop computers. Passive heat sinks on the other hand are cooled by air flow across the heat-sink fins. The air flow is usually provided by one or more system fans and sometimes may be ducted from the fan face to the heat sink. Passive heat sinks with or without ducted air flow are used widely in workstations and servers.

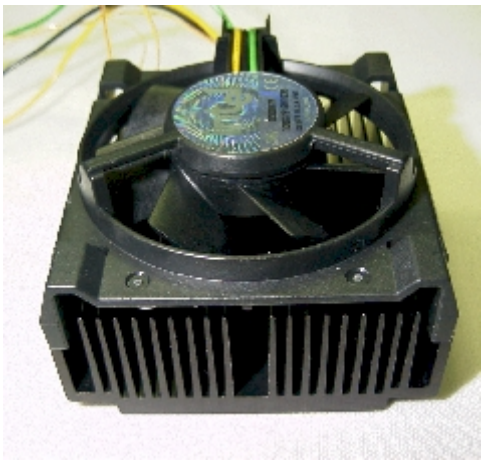


Figure 22: An active heat sink

In addition, heat sinks are also classified based on the method of manufacturing as follows:

- extruded heat sinks
- folded-fin heat sinks
- integrated vapor-chamber heat sinks

These heat sink types are discussed in the following sections.

Extruded Heat Sinks

Extruded heat sinks are usually made of aluminum and are manufactured by extruding a large billet of material through a die to provide the fin shape. Such an operation usually results in a heat sink with rectangular longitudinal fins like the one shown in Figure 17. The exact shape of the fin is rarely rectangular with a fin thickness that is larger at the base than at the fin tip. Pin-fin heat sinks (see Figure 18) are manufactured by using a cross-cut operation with a milling machine. In most cases, the extrusion and or machining process is followed by an anodization step that produces the black or colorful heat sinks often seen in desktop computers.

Extruded or machined heat sinks are expected to provide a convective resistance in the order of 1.0-1.5 °C/W for the typical air flow rates available in desktop computers. The extrusion process places a limitation on the fin height to fin gap that can be manufactured, primarily resulting from considerations of the structural strength of the die used to extrude the fin shape. Typically this ratio is of the order of 8:1. Extrusion is a highly automated, high-volume process and offers significant savings in manufacturing cost. If fins at a smaller pitch are desired, other manufacturing processes such as machining or die casting must be used. Although die casting and machining can provide heat sinks with a denser fin array, the manufacturing costs are also higher.

Folded-Fin Heat Sinks

The higher power dissipation of processors require a low cost, automated manufacturing process that can deliver fin arrays with a pitch much tighter than that available with the traditional machining and die-casting processes. This has been achieved through the use of folded-fin technology (see Figure 23) where the ratio of the fin thickness to fin pitch can be as low as 1:3. In this process, the fins are formed by bending (or folding) a strip of sheet metal aluminum or copper into an array of fins. The fin array is then bonded to a heat sink base made of aluminum or copper. Copper folded fins can be brazed or soldered to the copper base, thereby eliminating the fin-bonding resistance. Typical high-volume manufacturing processes for aluminum fins utilize epoxy bonding, which may introduce an additional fin-bonding resistance in the order of 2°C cm^2/W. Recent developments in manufacturing have utilized nickel plating or copper-flash treatment on aluminum fins to allow the use of brazing. This eliminates the fin-bonding resistances, resulting in thermal performance parity with copper-fin heat sinks, albeit at a lower weight.

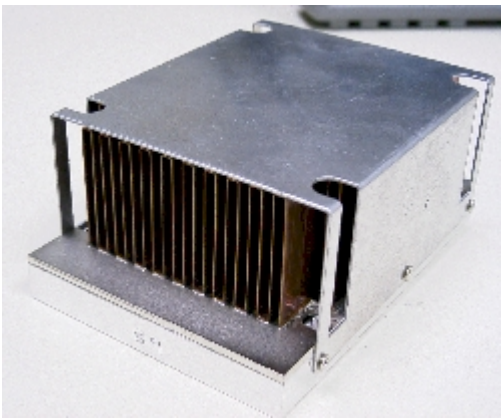


Figure 23: Folded-fin heat sink with a shroud

Typical heat sink thermal resistances obtained using folded-fin heat sinks are of the order of 0.3 to 0.6 °C/W at 15 to 20 cfm of dedicated air flow. This is a nearly twofold improvement over the performance of extruded-fin heat sinks.

Integrated Vapor-Chamber Heat Sinks

The resistance to heat spreading is primarily governed by the thermal conductivity of the heat sink material. One way to reduce spreading resistance is through the use of a heat pipe. Figure 24 shows a schematic of a typical cylindrical vapor chamber that consists of an evaporator, an adiabatic section, and a condenser [7]. Fluid vaporizes in the evaporator and condenses in the condenser section. In an actual application, the evaporator is placed in

contact with the processor, and the condenser is cooled by forced convection. Since the evaporation and condensation temperatures are identical, an ideal heat pipe is expected to move heat from the hot to the cold regions with negligible temperature drops.

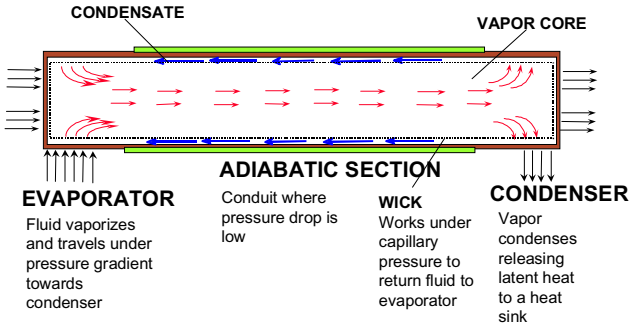


Figure 24: Schematic of a vapor chamber or heat pipe

A novel technique for improving the heat-sink performance is to judiciously combine vapor-chamber and folded-fin technologies [8]. The heat-sink design consists of a hollow vapor-chamber base that functions like a heat pipe. Folded fins are bonded to the top of the vapor chamber to form a heat sink. Typical heat-sink thermal resistances of 0.2 to 0.4 °C/W can be expected using vapor-chamber-folded-fin heat sinks at an air flow rate of 15 to 20 cfm. However, the cost of manufacturing vapor chambers is nearly five to ten times that of extruded fin heat sinks. Figure 25 shows an example of thermal performance for the various heat sink technologies that are used in the desktop and server market segments.

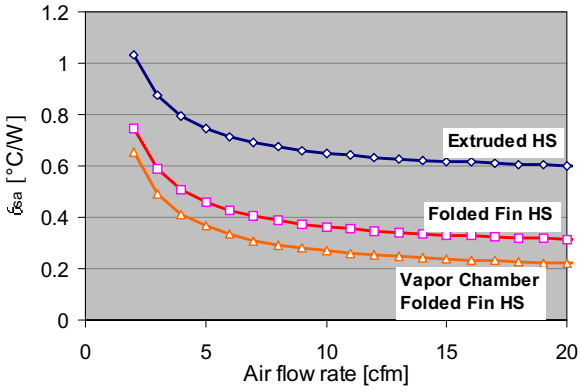


Figure 25: Thermal performance of desktop and server heat-sink designs

Figure 26 shows a summary of heat sink cost versus performance. It can be inferred that to meet the demands of higher power processors, increasingly complex heat sink technologies need to be deployed (lower θ_{sa},). This leads to a corresponding increase in the unit cost of the heat sink. Thus,a system designer must do a cost-performance analysis and select an optimum cooling solution based on the geometric, cost, and weight-boundary conditions for a given market segment.

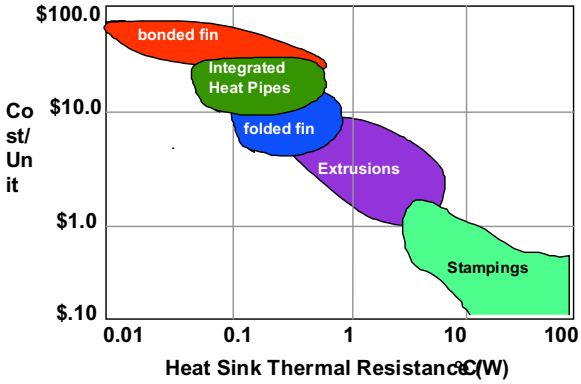


Figure 26: Heat-sink cost vs. performance

Figure 27 shows a thermal technology map for the various market segments. It shows the heat-sink volumes and thermal-resistance requirements for the various system platforms. The figure shows that the available space for heat sinks is roughly 20 to 30 cubic inches for the desktop and workstation-server platforms.

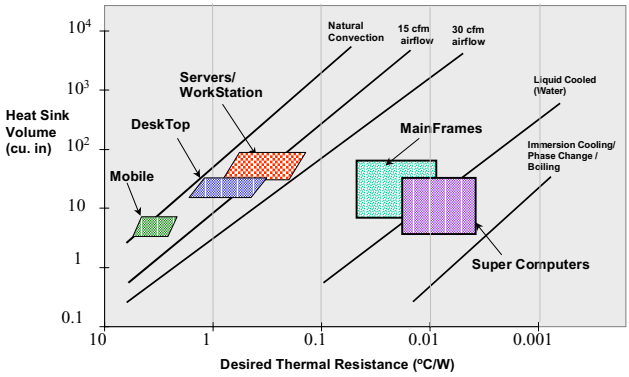


Figure 27: Thermal technology map

The volume available for cooling determines the geometry of the heat sink. Thus, to obtain better thermal performance, the fin spacing on the heat sink must decrease. Additionally, heat spreading in the heat sink must be improved to reduce the heat sink thermal resistance. Both of these requirements lead to more expensive heat-sink technologies like folded-fins and vapor-chamber heat-sink bases. The heat-sink volume available space in mobile platforms is even smaller, around 8-10 cubic inches, necessitating the use of heat

pipes for remote cooling. Note also that next-generation system designs for all these market segments are demanding lower heat-sink thermal resistances in combination with shrinking heat-sink volumes. These requirements make the task of thermal management for the system designer even more challenging.

CONCLUSION

The need for an integrated approach to deal with the thermal challenges posed by next-generation processors is clear. Improvements in one single area alone will not be able to satisfy the thermal-budget requirements. Architecture and design techniques, process shrinks, and voltage scaling are critical to maintain a manageable power-frequency roadmap. Focus needs to continue on packaging materials and technologies to reduce interfacial resistance and improve heat spreading. Board layout designers need to pay attention to the layout of high- and medium-power devices in the vicinity of the microprocessor. System designers need to focus on optimizing air flow and preheating from other components in the chassis. The desired outcome would be to drive design and technology development concurrently at silicon, package, motherboard, and system-level packaging to ensure that thermal solutions can support the demand for increasing computing and communication needs.

ACKNOWLEDGMENTS

The authors thank Chia-Pin Chiu, Gary Solbrekken, and Hong Xie for some of the graphics used in this paper.

REFERENCES

[1] Solbrekken, G. L., Chiu, C. P., Byers, B. and Reichenbacher, D., "The Development of a Tool to Predict Package-Level Thermal-Interface Material Performance," *Proc. 7th Intersociety Conference on Thermal and Thermo-Mechanical Phenomena in Electronic Systems*, Vol. 1, 2000, pp. 48-54.

[2] Incropera, F. P. and DeWitt, D. P., *Fundamentals of Heat and Mass Transfer*, 4th Edition, John Wiley & Sons Inc., New York, 1996.

[3] Kraus, A. D. and Bar-Cohen, A., *Design and Analysis of Heat Sinks,* John Wiley and Sons Inc., New York, 1995.

[4] Lee, S., "Optimum Design and Selection of Heat Sinks," *IEEE Transactions on Components Packaging and Manufacturing Technology Part A*, Vol. 18, No. 4, 1995, pp. 812-817.

[5] Xie, H., Aghazadeh, M., Liu, W., and Haley, K., "Thermal Solutions to Pentium Processors in TCP in Notebooks and Sub-Notebooks," *IEEE Transactions on Components Packaging and Manufacturing Technology Part A*, Vol. 19, No. 1, 1996, pp. 54-65.

[6] Viswanath, R and Ali, I. A., "Thermal Modeling of High-Performance Packages in Portable Computers," *IEEE Transactions on Components Packaging and Manufacturing Technology Part A*, Vol. 20, No. 2, 1997, pp. 230-240.

[7] Faghri, A., *Heat Pipe Science and Technology*, Taylor and Francis Publishers, 1995.

[8] Yusuf, I., Watwe, A., and Ekhlassi, H., "Integrated Heat Sink Heat-Pipe Thermal Cooling Device," *Proc. 7th Intersociety Conference on Thermal and Thermo-Mechanical Phenomena in Electronic Systems*, Vol. II, 2000, pp. 27-30.

AUTHORS' BIOGRAPHIES

Ram Viswanath graduated from Rutgers University with a Ph.D. in Mechanical and Aerospace Engineering. He then joined Intel in 1993 and works with the Assembly Technology Development Group in Chandler, Arizona. He has authored numerous technical papers in refereed journals and conferences. His current focus is on package design and performance optimization for 0.13µ microprocessors. He holds multiple patents in the areas of thermal management tools and techniques for assembly and testability of microprocessors. His e-mail address is ram.s.viswanath@intel.com

Vijay Wakharkar graduated with a Ph.D. degree in Materials Science and Engineering from SUNY at StonyBrook in 1989. He is currently managing the materials group responsible for polymers and heat spreader materials and supplier development within the Assembly Technology Development Group. Vijay has worked at Intel for nine years on materials development projects supporting the various package technology efforts within ATD ranging from TCP, PPGA, PLGA, Cartridge (SECc), and Flip-Chip Technology. Prior to working at Intel, Vijay spent two years as a Post Doctoral Associate at the IBM Alamden Research Center in San Jose. His e-mail is vijay.s.wakharkar@intel.com

Abhay Watwe obtained his M.S. degree in Mechanical Engineering from the University of Houston and his Ph.D. in mechanical engineering from the University of Minnesota in 1996. He worked with Fluent Inc. as an applications engineer and joined Intel in 1997 as Senior Mechanical Engineer. Abhay has published roughly 20 papers in archival journals and conferences during the last five years. He is also coauthoring a chapter on thermal management of electronics in a new text book for

undergraduate students to be published in October 2000. His e-mail is abhay.watwe@intel.com

Vassou LeBonheur obtained a B.S. degree in Chemical Engineering from the Ecole Superieure de Chimie Industrielle de Lyon, France and a Ph.D. in materials science and engineering from the University of Illinois at Urbana-Champaign. He joined Intel right out of college. He is a Senior Materials Engineer in the Materials Enabling & Technology group in Assembly Technology Development. Vassou has been at Intel for four years and has worked on the development of different materials (encapsulant, die-attach, thermal-interface materials, sealant and adhesive, thermal plates, etc.) in support of several development programs including PLGA, Cartridge (SECc), and Flip Chip. His e-mail is vassou.lebonheur@intel.com