Preface

Lin Chao
Editor
Intel Technology Journal

Microprocessor packages are the external "suits" for the complex and intricate world of silicon chips. The working part of a microprocessor is a small piece of silicon no larger than a postage stamp encased in a sealed "package." The chip itself must be sealed away to prevent external contaminants, such as dust, from adversely affecting the silicon chip. The art and science of semiconductor packaging has advanced radically over the last decade as faster and more powerful microprocessors with millions of transistors stressed the state of the art in microprocessor packaging. Thermal-heat dissipation, signal interconnects, and higher densities have required many advances. Pin-grid arrays with hundreds of pins, multicavity modules, leadless chip carrier and quad flat packs are the many types of microprocessor packages today.

The seven papers here present an engaging discussion on Intel's microprocessor packaging technologies. They highlight the technical challenges faced by packaging developers now and in the future, and in a broad sense, ties them into the many challenges faced by the semiconductor industry to achieve the next level of performance. The first paper traces the evolution of Intel's microprocessor packaging technologies. Flip-Chip Pin Grid Array (FCPGA) used in Intel's high-performance microprocessors uses balls of solder and gold that are melted (or reflowed) to connect the silicon chip to the package. The second and fifth papers look at this packaging technology.

The third paper explains the technical complexity of interconnect design to achieve optimal electrical performance. This paper discusses the design analysis and synthesis techniques used to ensure optimal electrical design. The fourth paper presents the challenges faced in thermal design. Ensuring that packaging continues to meet high standards of reliability is a key to success and is discussed in the sixth paper. Finally, the seventh paper discusses the practical problem of managing the thermal environment during microprocessor testing.

Microprocessor Packaging: Evolution and Future Challenges

By Nasser Grayeli
Director, Assembly Technology Development
Intel Corp.

The past decade has seen the evolution of microprocessor packaging from a simple protective scheme to a complex combination of different elements that enable microprocessor performance while still providing the basic function of protection. Packaging today's microprocessor on the one hand entails tailoring the package to enable microprocessor performance, a complex task considering the rapid rate of microprocessor performance growth. This challenge is in terms of schedule and technical complexity. On the other hand, the package forms the interface between the microprocessor and the external world of the motherboard and the computing system. In this capacity, package design must allow for an easy interface and must meet a diverse set of form factor requirements.

The package provides a conduit for the microprocessor through a space transformation allowing small-scale features on the silicon to be electrically connected to the external environment. This is a challenging geometrical problem and requires that packaging interconnection densities must closely track the evolution of microprocessor interconnection densities. In connecting the die to the motherboard, the package must also ensure that the connections do not unduly inhibit the microprocessor performance by introducing unnecessary electrical impediments usually referred to as package “parasitics.” As microprocessors have evolved, they have increased in speed, which in turn needs increasingly sophisticated power delivery schemes. Another consequence of microprocessor evolution has been increasing power dissipation. Package design must now provide a path for thermal dissipation, requiring a better understanding of the thermal characteristics of packaging materials and design. Package design also requires a good understanding of the structural characteristics of the package to ensure it is designed for reliability and robustness. Attention is increasingly focused today on understanding the electrical, thermal and mechanical characteristics of packaging to optimize all these aspects.

The package is also the interface that connects the microprocessor to the motherboard. In this capacity it must have a compatible interface to allow for easy acceptance on the motherboard as well as the system design. The form factor of the package is a critical element for easy interface to the motherboard. The requirements are usually different in different market segments and often drive the need for form factors that are tailored to these different segments. For instance, the height of the package is critical to enable a microprocessor in a mobile market where a slim and low weight package is critical to success. On the other hand, the ability to dissipate high power, and hence features that enables this, are critical in a server or desktop market segment. Cost, compatibility and fit within the computer system are key parameters that must be designed for in making a microprocessor successful. This challenges us into concurrently developing

multiple solutions and technologies geared towards specific market segments.

Aside from the challenges of package design, there is a need to develop efficient and cost-effective manufacturing processes that allow us to meet the schedule and volume demands of today's market places. These have presented us with interesting challenges in understanding the manufacturability, testability and reliability of packaging. Some of these issues are discussed in greater detail in this issue.

This Q3'00 issue of the Intel Technology Journal has been designed to provide the reader with a broad scope view of the technical challenges in the design, manufacturing, testing and reliability assessments of microelectronic packaging. By focusing on microprocessor packaging, which represents the technical envelope and the greatest challenges, the papers in this volume attempt to highlight different aspects of the evolution and future of packaging. A perusal of this journal will help the reader better appreciate the systematic manner in which we have successfully addressed the challenges of today and how we continue to plan for the future.

Flip-Chip Technology on Organic Pin Grid Array Packages

Mirng-Ji Lii, Assembly Technology Development, Intel Corp.
Bob Sankman, Assembly Technology Development, Intel Corp.
Hamid Azimi, Assembly Technology Development, Intel Corp.
Hwai Peng Yeoh, Assembly Technology Development-M, Intel Corp.
Yuejin Guo, Assembly Technology Development, Intel Corp.

Index words: flip chip, organic, pin grid array, surface mounted technology pin

ABSTRACT

As microelectronic devices become more integrated with increased functionality and higher levels of performance, the complexity of packaging technology grows proportionally. Today's silicon processes have enabled microprocessor designs to achieve very high clock frequencies. As a result of the increase in feature integration, high clock frequencies, and the power supply requirements of the latest generation of microprocessors, the density of interconnects between processor chip and substrate has been increased remarkably. New package substrate technologies with enhanced interconnect density are required in order to take full advantage of these silicon advancements. This has created an array of challenges in package design, substrate technology development, and assembly processes development. To provide a highly integrated and lower cost package, the Flip Chip Pin Grid Array (FCPGA) package was proposed as an innovative packaging solution [1]. This package utilizes laser-drilled blind/buried vias stacked on a PTH to ease routing and to lower the power supply loop inductance. In addition, the integration of flip-chip technology on an organic substrate helps to provide adequate signal and power supply interconnects. The FCPGA package was designed as a socketable solution. By taking advantage of the existing PGA socket infrastructure, this package helped to expedite the Original Equipment Manufacturers (OEMs) acceptance of the new package. This paper also describes the challenges encountered in the past in package design, validation, and assembly process development. Several technical challenges such as meeting the stringent impedance requirement to enable RDRAM* bus functionality, the optimal pinning process to certify Surface Mounted Technology (SMT) pins, and Underfill material and process development to fulfill throughput and performance requirements were overcome. The FCPGA package not only delivered a package with high performance on a cost-effective substrate, but also intelligently reused existing assembly equipment to minimize overall packaging cost. With the success of the first-generation FCPGA package technology certification, which has been utilized in the Intel® Pentium® III microprocessors, future generations of this technology will be developed that should offer great advantages for future Intel products.

INTRODUCTION

The need for high-density interconnects in a cost-effective flip-chip package was the motivation for FCPGA technology development. This paper describes the challenges encountered during the first generation FCPGA package design, validation, assembly processes, and material development.

FCPGA was designed as a socketable solution. The pin side view of an FCPGA package is shown in Figure 1. The use of the existing 370 socket infrastructure helped with the OEM acceptance of this new package.

The key features of the FCPGA technology are as follows:

1. Stackup

The substrate is comprised of an FR-5 equivalent core with two resin build-up layers on each side. Both blind and buried vias are used to ease package routing.

2. Bump Pitch

The flip-chip interconnects are built on an organic substrate with a solder bump pitch of 11 mils (279.4 μm).

* Other brands and names are the property of their respective owners.

3. Decoupling Capacitors

Pin side decoupling capacitors were added to lower the power supply loop inductance.

4. Surface Mount Package Pins

SMT pins were used to ease package routing. This was an improvement over through hole mounted pins. The use of SnSb solder to join the package and pins provided solder joint reliability through subsequent reflow operations.

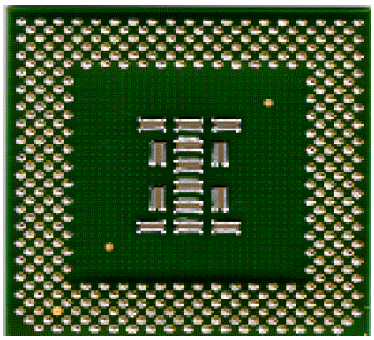


Figure 1: Pin side view of the FCPGA package

Package Design and Validation Overview

Package Designs

Several test vehicles and test structures were designed and analyzed to validate the package's electrical, thermal, mechanical, and reliability performance. Key attributes of several packages are tabulated in Table 1 [2, 3, 4, 5].

Table 1: Package design attributes

Attributes	Test Package A	Test Package B	Test Package C
Form Factor	1.95" x 1.95"	1.95" x 1.95"	1.95" x 1.95"
Thickness	1.1 +/- 0.1 mm	1.1 +/- 0.1 mm	1.1 +/- 0.1 mm
Package Layers	6 layers	6 layers	6 layers
Mn. Bump Pitch	279 µm	279 µm	279 µm
Bump Pattern	FCR in three I/O rows Square gird in core area	FCR in three I/O rows Square grid in core area	FCR in three I/O rows Slight offset parallelogram grid in core area
# of C4 bumps	1199	1286	1209
Die Size	0.355" x 0.455"	0.440" x 0.363"	0.438" x 0.386"
Die Layers	short loop	full loop	full loop
Core Voltage	>25V	1.5V / 1.6 V	1.55V / 1.8V
Package Stackup	L1/L2 : signal layers L3-L6 : plane layers	L1 : signal layers L2, L5 : partial signal layers L3, L4, L6 - plane layers	L1, L2 : signal layers L3-L6 : plane layers
Vias	Single-layer µ-vias Two-layer µ-vias	Single-layer µ-vias	Single-layer µ-vias
Footprint	PGA_370	PGA_370	PGA_370
# of chip cap.	18	14	7
Power Dissipation	>30 W	15~28 W	15~20 W

The layer structure of an FCPGA substrate is displayed in Figure 2; the targeted thickness of each layer and package key feature sizes are given in Table 2.

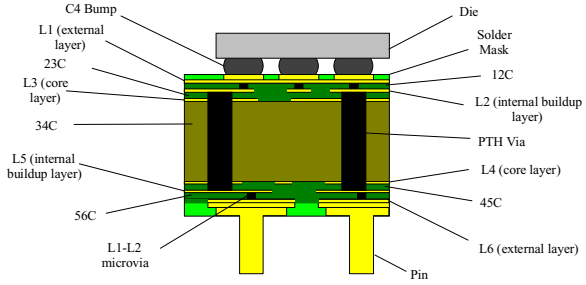


Figure 2: Cross section of FCPGA substrate

Table 2: Mean thickness of FCPGA stack up

Label	Feature	Thickness (SI)
	Solder Resist over Copper	25 µm
12C, 56C	External Buildup Layer Dielectric	30 µm
23C, 45C	Internal Buildup Layer Dielectric	30 µm
34C	Core Layer Dielectric	800 µm
L1, L6	External Buildup Layer Copper	17 µm
L2, L5	Internal Buildup Layer Copper	25 µm
	Copper over PTH	17 µm
L3, L4	Core Layer Copper	14 µm
	Total Package Thickness	1.1 mm

Electrical Characteristics

Empirical measurements and electrical modeling were used to assess the characteristic impedance (Zo), inductance, capacitance, resistance, and dielectric of the package. These parameters will impact the overall design by influencing the signal integrity, power supply droop, and routing requirements [6].

One of the resistance test structures built into the test packages is shown in Figure 3; four point probing was used for the resistance measurement. Multiple via chains and conductor sheet resistance data confirmed that the package manufacturing process was capable of meeting targeted specifications. In order to ensure impedance values satisfied the data bus requirements, dielectric constants across a wide range of frequencies were analyzed and measured. The comparison between modeled and measured values is shown in Figure 4. The empirical data is in good agreement with analytical prediction.

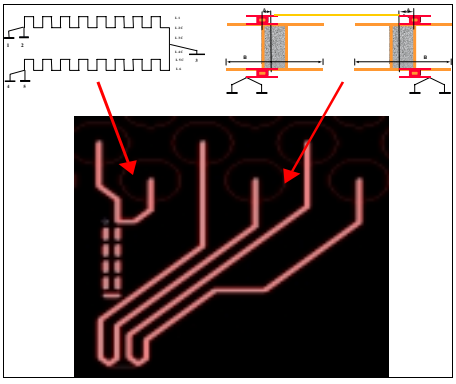


Figure 3: FCPGA resistance test structure

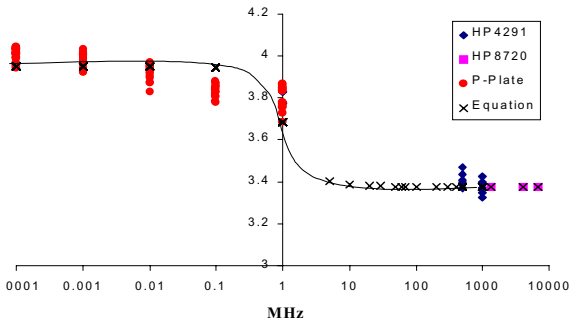


Figure 4: Comparison of calculated and measured dielectric constants in various frequency ranges

Thermal Performance

Both modeling and testing were conducted to validate the FCPGA thermal solutions. With increasing core speed and maximum power dissipation, both passive and active heat sinks were evaluated. Details of the thermal design challenges are discussed in the Thermal Designs section of this paper.

Mechanical and Package Reliability

Mechanical tests and modeling were performed to address concerns about the structural integrity of the FCPGA package. A total of 100 FCPGA samples were tested with uniform and edge-loaded forces (20 to 100 lbs). Visual inspection and post-stress electrical test data confirmed that there was no change in the mechanical and electrical integrity of the package. Moreover, an additional 40 samples were uniformly loaded up to 100 lbs. and subjected to 600 cycles of T/C "B." No sign of failures was seen, and the chip cap solder joint strength retained a healthy level as illustrated in Figure 5.

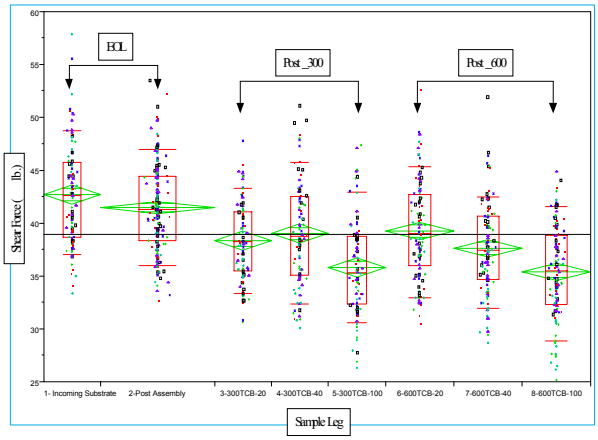


Figure 5: Chip capacitor solder joint shear strength distribution at end-of-line, 300 and 600 cycles of T/C "B"

The collected data on maximum package/die loading and chip cap solder joint shear strength confirmed the

robustness of the FCPGA package. Various stresses specified for Intel assembly technology certification (such as Temperature Cycling, Bake, Power Cycling, Shock, and Vibration, etc.) were performed to accelerate other possible failure mechanisms. Several failure modes (such as metal migration and weak pin solder joint) were observed early on, but fixes were quickly implemented, which eliminated these issues. In conclusion, there were no high-risk issues that appeared during testing that impacted the technology certification.

Continuous data was collected at Intel and at supplier manufacturing sites. This included electrical, mechanical, and thermal measurements. These data, collected since early in the development phase, built sufficient confidence that the FCPGA package was a viable packaging solution for current as well as future microprocessors.

SUBSTRATE DEVELOPMENT OVERVIEW

Photolithography and etch have been the most prevalent methods to create blind µ-vias in high-density substrates. The photolithography process has two main disadvantages. First there is the limitation to the µ-via diameter due to the limited resolution of commercial photosensitive materials. Second, photosensitive materials are prone to reliability issues with their mechanical properties, moisture absorption, and the value of the dielectric constant. Laser µ-via drilled into an "off the shelf" dielectric can overcome these limitations.

Laser technology can potentially create via sizes down to the < 10 µm range, while today's photolithography materials are limited to 50-60 µm vias. Another plus is that the smaller the via size, the lower the cost of the laser µ-via formation due to a shorter pulse time. In addition, by eliminating photosensitive resins, a large number of non-photo sensitive materials can be considered for dielectric material. The laminate material used in FCPGA is a commercially available film, which is lower in cost when compared to the dielectric materials used in other photo µ-via based packages.

In the FCPGA package, the laser drill via technology was implemented in spite of the fact that the line/space design rules were not as challenging as the existing Organic Land Grid Array (OLGA) technology. The rationale was to save money on processing costs, while taking advantage of the higher routing density resulting from the smaller µ-via pads. By using laser vias instead of photo vias, the FCPGA package had access to cheaper and better commercially available dielectric materials. Figure 6 shows a schematic of the laminates and materials used in the FCPGA.

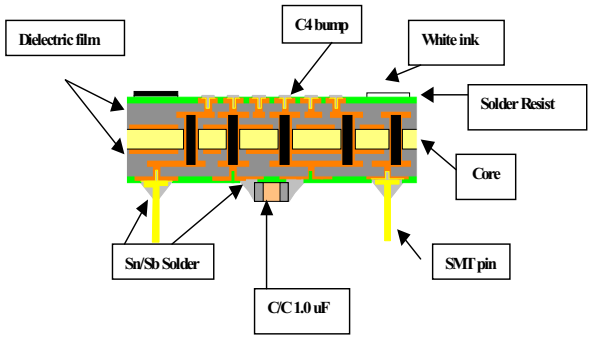


Figure 6: Schematic of FCPGA laminates and materials

Comparing FCPGA with the previous OLGA package, there are several distinct differences:

1. FCPGA uses laser vias instead of photo vias, and SMT pins.
2. The size of the FCPGA package is larger with additional area for chip caps.
3. Commercially available dielectric and solder-resist materials are used in the FCPGA package.
4. µ-via is used on PTH in the FCPGA package.

FCPGA KEY CHALLENGES

Thermal Designs

Thermal design solutions for the FCPGA package pose challenges because of the system chassis spatial constraints and the need to meet maximum power dissipation requirements. The design was also challenging because a heat sink ground feature that suppresses potential electromagnetic emission had to be integrated into the package. The key thermal design constraints are listed in Table 3. To broaden the variety of possible FCPGA package applications, both passive and active heat sink designs were evaluated. A schematic of passive and active heat sink solutions is shown in Figures 7 and 8. Preliminary empirical and modeling results suggested that passive and active solutions could support power of about 19W and 22W, respectively.

Table 3: Summary of typical FCPGA thermal design attributes

Attributes	Product A	Product B
Thermal Design Target	19.3 W	22 W
Tj	90 oC	85 oC
Ta (system internal)	45 oC	45 oC
Theta_ja	2.33 oC/W	1.59 oC/W
Airflow = 200 fpm	200 fpm	150 fpm
Clip force requirement	12 - 20 lb	12 - 20 lb
HS design	passive	active
HS weight	140 - 180 g	140 g

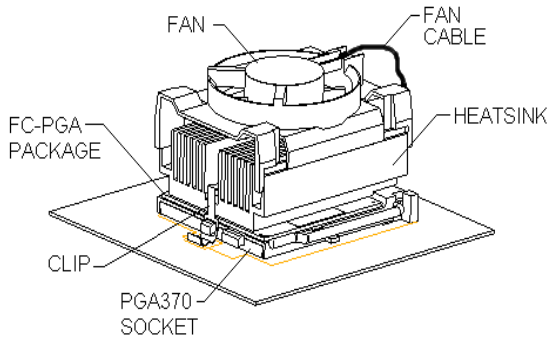


Figure 7: Schematic of an active heat sink solution

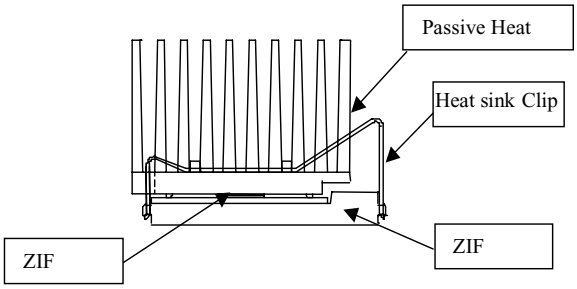


Figure 8: Schematic of a passive heat sink solution

High-Speed Bus Impedance Requirement

High-speed digital systems have problems that manifest themselves in three distinctive ways. First, conductor traces can experience reflections due to multiple impedance mismatches. Second, cross talk may occur because of unwanted electromagnetic coupling between adjacent traces. Third, ground bounce can be significant due to inductance in the ground return path of the IC package. The combined result of these effects could adversely impact timing margins in systems and thus limit the ultimate performance of the system. In the FCPGA package, a good deal of effort was put into controlling the impedance to limit the impact of the first problem.

The design of a fixed-impedance bus structure makes it more sensitive to the physical dimension tolerances in the manufacture of the package. The FCPGA design rules will support multiple impedance targets in the package. The bus impedance specifications call for a tolerance target of +/- 10%.

Impedance is a function of dielectric layer thickness, dielectric constant, and Cu trace width. This requires rigorous tolerance control of dielectric thickness. Non-uniformity in thickness of the Cu plating will directly add variability to the thickness of the dielectric layer between two adjacent Cu layers. Since there is inherent Cu thickness variation, the control of dielectric layer thickness variation becomes even more stringent.

Initial data indicated this variation would be a challenge for the FCPGA package as there was a higher dielectric thickness variance than desired, and a correspondingly variable impedance value. However, as illustrated in Figure 9, impedance (Zo) measurements taken after process improvements showed that the mean impedance stabilized. The improvement was made possible through improving the Cu thickness uniformity and by making a smoother insulator surface.

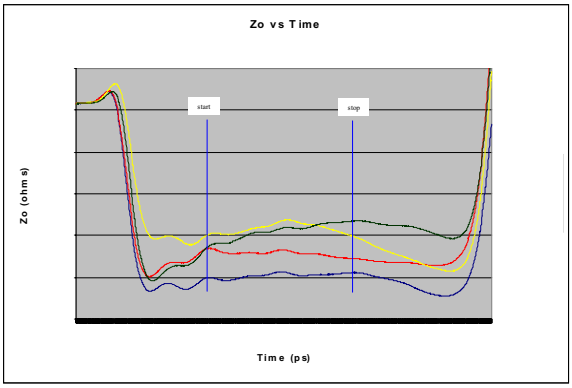


Figure 9: Impedance measurements with two different dielectric thicknesses

µ-Via Reliability Issues

Preliminary reliability data showed µ-via delamination. The µ-via delamination resulted in a high-percentage fall-off after 300 cycles of T/C "B" (-55C <-> 125C) stressing and higher cumulative fails after 1000 cycles of T/C "B". Figure 10 shows an example of a delaminated via that caused an electrical failure.

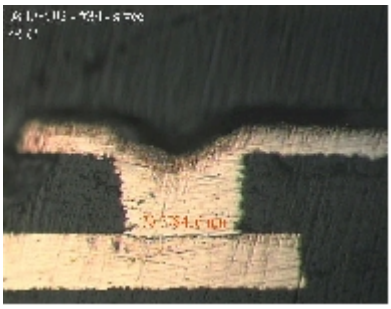


Figure 10: Example of a delaminated via

Failure analysis suggested that via delamination occurred at the interface between the electrolytic Cu and electroless Cu layers. Hot oil thermal shock tests were utilized as a quick-turn reliability monitor. The measurement of improvements in the manufacturing process were based on the shifts in the µ-via resistance. Test data confirmed that two factors were the significant modulators in µ-via delamination. Figure 11 shows a schematic drawing of various via structures, as well as the test data for µ-via resistance shift after hot oil thermal shock. The test results clearly indicated that via resistance increased more than 50% after thermal shock for some test structures. After process fixes were implemented, no via resistance shift was seen after 400 cycles of stress!

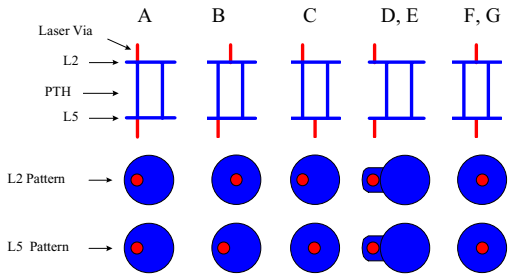


Figure 11a: Schematic of various via structures

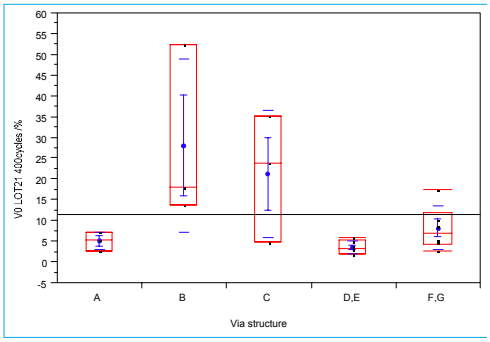


Figure 11b: Via resistance before process fix

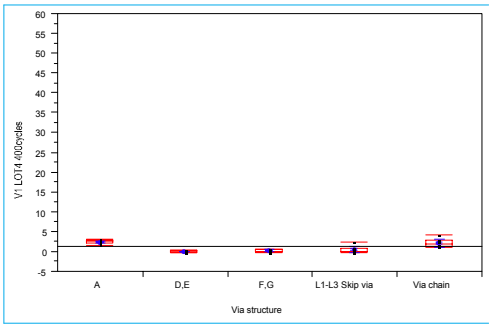


Figure 11c: Via resistance after process fix

The improvement in via integrity was also verified through via "pop" tests. Three types of failure modes were observed and are shown in Figure 12. None of these failure modes appeared on FCPGA packages after the process fixes were implemented.

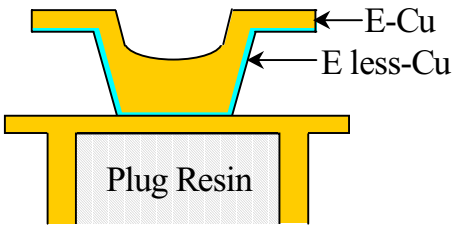


Figure 12a: Schematic of a PTH via structure

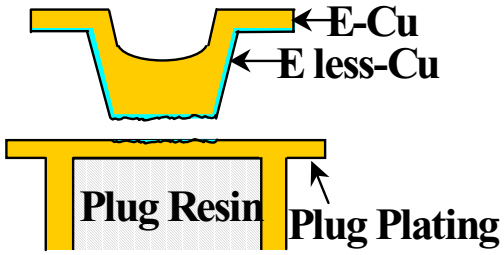


Figure 12b: Popped via failure indicating weak bonding at the two Cu layers interface on package before process fix

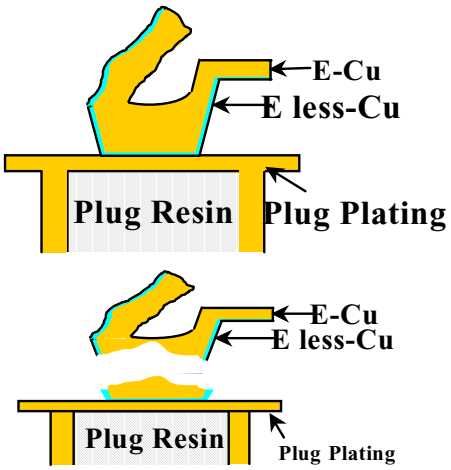


Figure 12c: Broken via edge indicating strong Cu layer interface on package after process fix

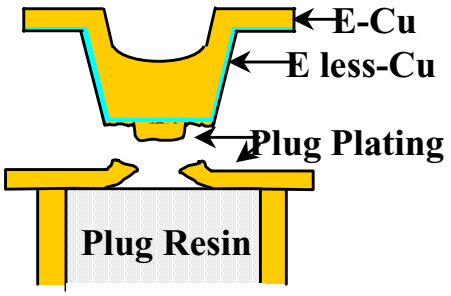


Figure 12d: Broken via plug plating indicating strong bonding at the two Cu layers interface after process fix

SMT Pin Development

The FCPGA package utilizes SMT butt-mounted pins on an organic substrate. To determine the reliability of these pins, experiments were conducted to evaluate different combinations of pin solder joint structures. They were made from three solder materials, three different solder volumes, skew misalign pins, and smaller pin nail-head sizes. The attributes of each leg of the experiment are detailed in Table 4. Additionally, pin pull and pin shear testing were used to quantify the pin strength before and after stresses.

Table 4 : SMT pin strength attributes

Legs	Nail Head Diameter (mm)	Solder Volume (mg)	Solder Composition	Pad Opening (mm)
A	0.9	M	SnAg	1.2
B	0.7	H	SnAg	1.2
C	0.9	H	SnAg	1.2
D	0.9	M	SnAg	1.2
E	0.9	L	SnAg	1.2
F	0.9	H	SnSb (A%)	1.3
G	0.9	H	SnSb (A%)	1.2
H	0.9	M	SnSb (A%)	1.2
I	0.9	H	SnSb (B%)	1.3
J	0.9	H	SnSb (B%)	1.2
K	0.9	M	SnSb (B%)	1.2

Test data confirmed that the smaller pin nail heads and intentional pin misalignment had lower pin strength as measured before assembly. Legs with M mg and H mg of SnAg and SnSb legs showed comparable pin pull strength and pin shear strength with the exception of the L mg SnAg lot, which had lower pin pull and pin shear strength. Interestingly, after assembly, the SnAg lots' pin pull strength was reduced significantly. Both SnSb (A%) and SnSb (B%) lots showed no sign of pin joint strength degradation after assembly. The SnSb (B%) was selected as the POR material because SnSb (A%) required a higher reflow temperature, which was undesirable.

Low Cost Underfill Material and Process Development [7]

The use of a two-step process and two separate materials for Underfill and fillet in C4-OLGA packages resulted in high equipment costs and a narrow process window. The challenge in FCPGA was to develop a low-cost, but high-performance Underfill material that would enable a simplified process and deliver high yields and improved unit per hour (UPH) capability.

Underfill material selection criteria included raw material cost, manufacturability, reliability and process integration performance, and supplier technical support and quality. The Underfill development team also reexamined the reliability and manufacturability success criteria such as alpha particle counts and the number of voids and voiding sizes, used in previous stages of development.

In developing the new Underfill process, viscosity and self-fillet formation are two key epoxy material properties. Based on material data sheets provided from fourteen suppliers worldwide, a total of four different materials was chosen for further evaluation. Score cards from each of the candidates were assessed to collect technical, business support, and quality data. Based on the collected information, the POR Underfill material was then selected. After the POR material was finalized, the

epoxy module engineering team focused on Underfill process optimization. Collaborative effort from the development team resulted in an optimized and simplified Underfill flow that met the FCPGA cost targets.

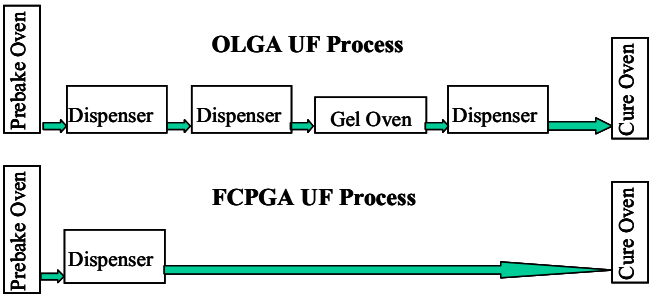


Figure 13: A comparison of the C4-OLGA and the FCPGA Underfill processes

Figure 13 illustrates the differences between the C4-OLGA and the FCPGA Underfill process flows. As shown in the POR flows, the FCPGA process has one, instead of two, dispensers and a BTU, which could save on equipment expenditures. Moreover, when results were evaluated, the FCPGA also improved the yield, and it had higher UPH throughput. The simplified Underfill process, together with the high performance of the Underfill material, was a plus for the FCPGA program.

Flip-Chip Solder Bump Non-Wet

Initial FCPGA data collection indicated the highest pareto of yield loss was attributed to open failures due to the non-wet of the C4 bumps. Low yield analysis revealed that the non-wet falls into two categories:

- edge non-wet, raccoon tail type, which can be detected with X-ray
- center non-wet, which is invisible with X-ray

However, in the second-phase data collection, a new oven was used for FCPGA reflow, and the open failures due to non-wet were reduced significantly. Low yield analysis confirmed the majority of non-wets were at the center of the die and invisible with X-ray.

A root cause assessment of the center non-wet showed no correlation between the microprocessor chip's passivation oxide thickness or co-planarity, the substrate's bump oxide, flux quantity, or uniformity during the chip attach process. A cross section of the center non-wet unit's solder bump joints revealed the substrate's and chips' bumps were not adequately aligned, except for those at the middle of the die. The right and left sides of the bump joints were slightly misaligned at opposite directions. Because of these observations, a failure mechanism for center non-wet was proposed; as illustrated in Figure 14, it turned out that the lack of bump co-planarity in the substrate prevented solder joints from forming at the center region of the die. This model also explained the slight shift in the alignment of the substrate's and die's bumps near the edge of the die.

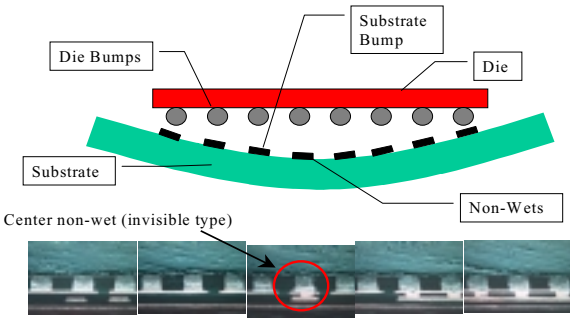


Figure 14: Proposed center non-wet mechanism; excessive substrate bump co-planarity prevents solder joints from forming at the center area of the die

The proposed model was validated through experimentation. From the data, we also realized that a maximum substrate bump co-planarity was required to prevent center non-wet. This has been defined in the specification for incoming packages.

FUTURE DEVELOPMENTS

The next generation of FCPGA package design rules have been defined and are currently under development. The performance enhancements in the new FCPGA2 package will be achieved with finer feature sizes (including bump pitches, Cu trace width and spacing, PTH size, stripline, etc.) and the addition of some new features (such as stacked vias and via-in-via). In addition, power delivery and removal capabilities will be improved through better decoupling capacitance and the use of a highly conductive package lid.

CONCLUSION

The FCPGA package design and development efforts have resulted in the integration of high-density flip-chip interconnect, SMT pins on an organic substrate. This new package is also high yielding, manufacturable, reliable, and low in cost. This cost-effective packaging technology represents a shift in direction from the previous OLGA technology, and it also represents a significant milestone in the evolution of organic substrate technology.

ACKNOWLEDGMENTS

The design and development of the FCPGA is a joint effort that spans several different organizations. We thank the entire FCPGA package design team, substrate development team, ATTD/QRE, and the assembly processes development staff for their hard work and devotion to support the development of this technology. We also acknowledge the great help received from the ATTD, ATTD/QRE/FA, and ATMO/CSMO divisions and other organizations. Finally, our special thanks to N. Grayeli, D. White, Jeff Watson, and M. Tay for their guidance and encouragement during packaging development. Finally, our thanks to the many others who contributed to the development of this program.

REFERENCES

[1] B. Sankman and M. Tay, "FCPGA Technology Target Specification," Intel internal document (1998).

[2] Jim Irvine and J. Dunham, "FCPGA Package-B1 Design Requirements Document," Intel internal document (1999).

[3] Jim Irvine and J. Dunham, "FCPGA Package-B2 Design Requirements Document," Intel internal document (1999).

[4] Tim Takeuchi, A. Waizman, A. Hasan and C. Baldwin, "FCPGA Package-C Design Requirements Document," Intel internal document (1999).

[5] C. Baldwin, "FCPGA Package-A Design Requirements Document," Intel internal document (1999).

[6] Rao Tummala and Eugene Rymaszewski, *Microelectronics Packaging Handbook*, (1988).

[7] Y. Guo, Y. Sha, C. Jayaram, and V. Wakharkar, "Low Cost Underfill Materials for Flip Chip Packages," *Intel Manufacturing for Excellence Conference (IMEC),* 2000.

AUTHORS' BIOGRAPHIES

Mirng-Ji Lii is a Packaging Design Integrator in the Assembly Technology Development group at Intel. He has various experience in microprocessor packaging technology development and interconnect development such as flip chip, tape-automated bonding, and fine pitch wire bonding. He holds two U.S. patents and has published several technical papers. Currently, he leads a package design team to develop a cost-effective, high-performance packaging technology for future-generation microprocessors. His e-mail address is mirng-ji.lii@intel.com.

Bob Sankman is the manager of the Design and Integration group in Assembly Technology and Development at Intel. He has held various technical positions in his career at Intel including wafer fab process engineer, failure analysis engineer, Q & R lab manager, and microprocessor package design manager. Bob has a B.S.Ch.E. degree from the University of Illinois. His e-mail address is bob.sankman@intel.com .

Hamid Azimi is currently the program manager for x60 substrate supplier certification in the Assembly Technology Development group at Intel. Hamid and his team work with Intel substrate suppliers to develop and certify a substrate material/process, which can meet cost, quality, reliability, and electrical performance targets. This includes substrates for mobile, DT, and server applications within the x60 generation. Hamid graduated with a Ph.D. degree in materials science from Leigh University in 1994 with expertise in fatigue and fracture of polymer composites and metals. He joined Intel in 1995 and since then has been working in the Assembly Technology Development group in the Material Technology Development Department for different platforms including PLGA, OLGA, FCPGA1 and, now, X60. His e-mail address is hamid.azimi@intel.com .

Hwai-Peng Yeoh is an Integration and Substrate Development Manager in the Assembly Technology Development group in Intel, Malaysia. He has more than eight years of experience in advanced microprocessor packaging technology development that includes ceramic packaging for the P5 generation, dual cavity ceramic packaging for the P6 generation, Plastic Land Grid Array packaging, Multichip Module packaging for Itanium, and Flip-Chip Pin Grid Array packaging. He currently leads both integration and substrate development teams in FC-BGA1 packaging development for high-performance chipsets. His e-mail address is hp.yeoh@intel.com .

Yuejin Guo has worked at Intel for years in the areas of assembly material development, new technology transfer, assembly process development, and assembly HVM production. His current position is as a Sr. Materials Engineer in the Assembly Technology Development group. Before joining Intel he was at Los Alamos National Lab as a researcher. Yuejin's technical interests are primarily in the area of packaging materials. He graduated from Caltech with a Ph.D. in chemistry. His e-mail address is yuejin.guo@intel.com .

The Evolution of Microprocessor Packaging

Ravi Mahajan, Assembly Technology Development, Intel Corp.
Ken Brown, Assembly Technology Development, Intel Corp.
Vasu Atluri, Assembly Technology Development, Intel Corp.

Index words: packaging, technology, building blocks, interconnect, power delivery/removal

ABSTRACT

Microprocessor packaging is undergoing major changes driven by technical, business, and economic factors. From the traditional role of a protective mechanical enclosure, the modern microprocessor package has been transformed into a sophisticated thermal and electrical management platform. Furthermore, microprocessor architecture and design techniques can have significant impact on the complexity and cost of packaging. The need to optimize the total solution (chip, package, board, and assembly) has never been more important to maximize microprocessor performance and minimize cost. It is important to point out that the package represents a way of connecting the microprocessor to the motherboard. In this capacity, it enables the fine feature, silicon-level interconnects to be connected to the motherboard, i.e., the package assists in a space transformation in a controlled and economically viable manner. The key to packaging is to ensure that it enables and optimizes microprocessor performance. In its early evolution, the influence of the package on microprocessor performance was limited; however, as the microprocessor evolves to provide increasing performance, the package must evolve to keep up, and packaging design must ensure that it optimally enables the microprocessor. Package performance, in this context, implies a clear understanding and optimization of the package's electrical, thermal, and mechanical characteristics to enable overall electrical performance and power dissipation and to ensure mechanical robustness. Recent advances in microprocessor packaging indicate a migration from wirebond (where the chip or die is interconnected to the package only on the periphery of the die) to flip-chip (where the die is interconnected to the package using the entire die area); and from ceramic to organic packages, with cartridge and multichip technologies emerging as key form-factors. With the emergence of the segmented market (mobile, desktop, server and associated sub-segments), we see a significant proliferation of packaging types tailoring functionality and costs to the different markets. To address this proliferation, Intel focuses on packaging building blocks that can be configured for different applications. This paper traces the evolution of Intel's microprocessor packaging technologies, delineates the technical and business drivers, and highlights emerging trends. It then highlights the technical challenges faced by packaging developers now and in the future, and in a broad sense, it ties them into the challenges highlighted in the semiconductor industry technology roadmaps. Finally, it provides an introduction to the other papers in this issue of the *Intel Technology Journal*, which deal in greater detail with some of the technical challenges discussed in this paper.

INTRODUCTION

In unit volume, microprocessors account for a small percentage (~ ≤ 1%) of the semiconductor components sold worldwide. However, their technical and economic significance are far greater. Microprocessor packaging represents the technology envelope of this discipline. To better understand this statement, we present a historical perspective of microprocessors and follow with a review of the motivators and technology directions for this component of the semiconductor industry.

THE EVOLUTION OF PACKAGING

In the Beginning: The Mechanical Enclosure

For many years, wirebonding and ceramic packages were the base assembly technologies for microprocessors because of their versatility and reliability. This was also true for Intel. Intel's 4004 microprocessor and later, the 8080, 8086, and 8088 microprocessors were all housed in ceramic dual-in-line packages (DIP) that used wirebond connections. By today's standards, these microprocessors had few Input/Output (I/O) pins (less than 40) and delivered very modest performance

(<20MHz). The primary function of the package was to provide space transformation (i.e., fan out) of the I/Os in order to ease board routing and protect the chip from mechanical damage and from the environment. These were simple, single-layer packages. Figure 1 shows a typical example.

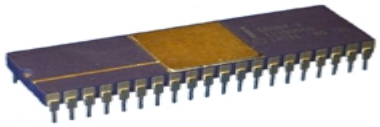


Figure 1: A 40 Lead DIP package used to package the 8088 and 8086 microprocessors

The Transition

In the i286™ and i386™ microprocessor generations, the number of I/O pins increased (~50 to 100) as greater functionality was incorporated into the microprocessor. This necessitated the use of Pin-Grid Array (PGA) packages in which a larger number of I/O connections could be accommodated in a small area. Also, in the i386 generation, it became evident that the increasing clock frequencies (a staggering 33MHz at the time) and simultaneous I/O switching could cause unwanted electrical coupling in the package manifesting itself as noise problems. Consequently, design and modeling competencies were substantially enhanced to account for these factors leading to the first use of multilayer ceramic packages at Intel. Figure 2 shows the i386 microprocessor package.

Figure 2: A 132L ceramic PGA package used for the i386™ microprocessor

Emergence of the Electrical/Thermal Platform

The i486 microprocessor was also housed in a PGA package with 168 leads. In addition to the basic function of connecting the I/Os, advanced electrical design concepts were incorporated. These included the use of power and ground planes as well as the inclusion of integrated capacitors in the package. These features transformed the package from a simple mechanical enclosure to a multilayer electrical distribution and signal-routing management platform.

The Intel® Pentium® processor helped in advancing this trend. In addition to the electrical features, the high-power dissipation, in the order of 15W, of the Pentium processor hastened the deployment of advanced thermal solutions such as an integrated heat spreader. These features, while effective, were expensive. An early version of the Pentium processor package is illustrated in Figure 3.

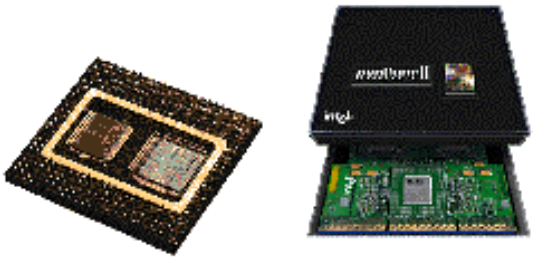
Figure 3: The Intel® Pentium® processor package in a ceramic PGA with a heat spreader

The Need for Increased Integration

The next-generation microarchitecture, commonly referred to as the P6 microarchitecture, introduced in the mid 1990s, represented a new era of performance and complexity. The microprocessor architecture called for a dedicated cache chip connected to the microprocessor via a Backside Bus (BSB). The first implementation of this architecture was on the Intel® Pentium® Pro processor where the microprocessor and cache chips were housed in a dual-cavity ceramic PGA package and connected by wire bonding. Because of the special requirement in the I/O configuration and because of the electrical performance of the cache memory of this microprocessor, custom Static Random Access Memories (SRAMs) were used, an expensive solution.

The second-generation implementation of the same microarchitecture utilized a cartridge form-factor. In this instance, the microprocessor and cache chips were housed in separate component packages and were connected using a standard printed circuit board. To start with, the microprocessor was assembled using Plastic Land Grid Array (PLGA) packages with wirebond technology, which later transitioned to Organic Land Grid Array (OLGA) packages that utilized Controlled Collapse Chip Connection (C4) technology. Aside from the transition from peripheral interconnect to area array interconnect, this packaging transition also marked the use of a high-performance package substrate technology,

i.e., OLGA technology. Plastic Quad Flat Package (PQFP) technology using wire bonding was used for the cache chips. This approach had two advantages over the dual-cavity ceramic PGA solution. First, it enabled the use of commodity Pipeline Burst SRAMs (PBSRAMs) thereby reducing cost. Second, the cartridge solution also allowed caches and other features to be customized for different market segments. The dual-cavity ceramic PGA and cartridge are shown in Figure 4. Figure 5 shows the portfolio of products packaged in a cartridge format.



(a) (b)

Figure 4: (a) First-generation implementation of the P6 microarchitecture in a dual-cavity ceramic PGA (b) The Single Edge Cartridge Connector (SECC) cartridge is a second-generation form-factor

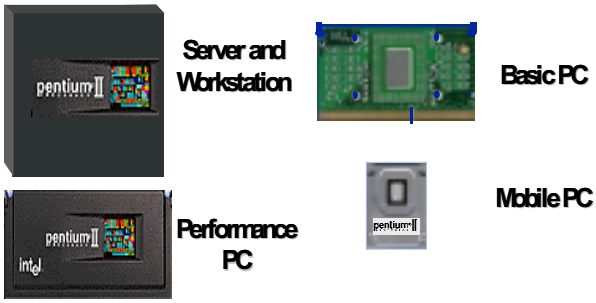


Figure 5: Portfolio of products utilizing cartridge packaging

Although the cartridge form-factor was an effective technical solution, the emergence of the cost-sensitive Personal Computer (PC) market demanded even more aggressive cost/performance packaging solutions.

Silicon Integration: Back to Single-Chip Packaging

Silicon feature scaling and the integration of the Level 2 (L2) cache directly into the microprocessor die were key enablers to lower the cost of packaging. Without the need for the multidie package or cartridge to service the high-speed BSB, it was possible to move back to single-chip packaging. Several single-chip packages were developed with form-factors based on market segmentation requirements.

Some of these microprocessor packaging form-factors included

- low-profile and high-density pinned packages for mobile applications (Figure 6)
- pinned packages for sockets in desktops (Figure 7); the package substrate is referred to as the Flip-Chip PGA substrate, another version of the organic substrate technology

(a)

(b)

Figure 6: Views of low-profile micro-PGA for mobile socket applications; the micro-PGA uses an OLGA substrate surface mounted to an interposer

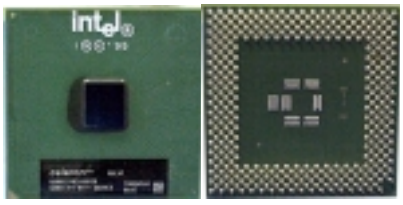


Figure 7: Pinned package for desktop socket; the package technology is referred to as Flip-Chip PGA (FC-PGA)

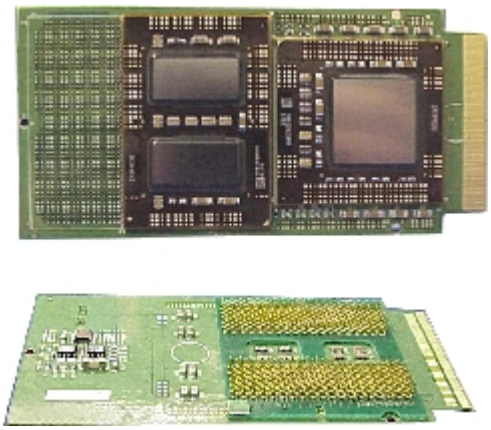


Figure 8: Intel® Itanium™ processor packaging shows how different elements of packaging can be combined

The focus at Intel has been to create packaging technology building blocks that can be combined to provide multiple features and form-factors, while minimizing piece part, process, and manufacturing costs. An example of this can be seen in Figure 8, which illustrates the packaging for the Intel® Itanium™ processor. However, before we discuss Intel's focus, we present an account of the technical and business drivers as well as the emerging directions for packaging technology.

TECHNICAL AND BUSINESS DRIVERS

Microprocessor packaging requirements are closely and intricately tied to the performance and architecture of the microprocessor. Figure 9 depicts the evolution of microprocessor/cache/bus architecture. Using this evolution as a framework, we examine five major drivers:

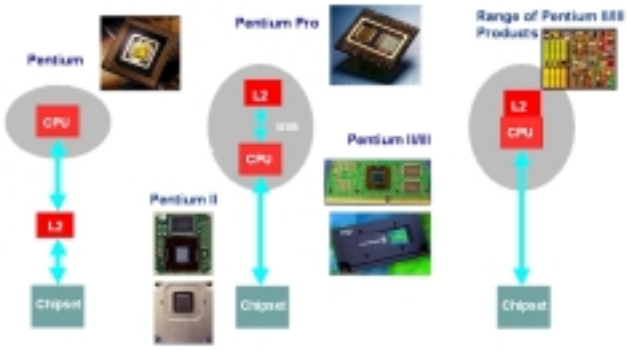


Figure 9: Evolution of the microprocessor and cache architecture from the i486™ microprocessor to the Intel® Pentium® Pro microprocessor

Driver #1 : Connecting the Cache

As the performance of microprocessors increased, the need to supply data and instructions to the microprocessor increased accordingly. This information normally resides in the main memories, such as the Dynamic Random Access Memory (DRAM) and disks, and it is channeled to the microprocessor via the bus, a parallel set of interconnects running between the microprocessor and the memory. The wider (i.e., more data lines) and faster the bus, the more data can be transferred at a given time. Since the days of the i486™ microprocessor, the speed requirements for data to be transferred to the microprocessor have exceeded the speed of the main memories (DRAMs). As a result, an L2 cache system utilizing fast Static RAMs (SRAMs) was added to the microprocessor architecture. This L2 cache stores frequently used data thereby reducing the need for frequent access to the external main memories. Consequently, this speeds up execution and leads to enhanced performance.

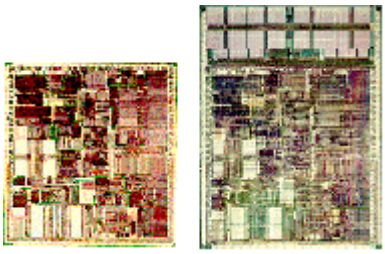
As described in the previous section, the Intel® Pentium® II/Pro microprocessor architecture had a dedicated BSB connecting the microprocessor to the L2 cache to further enhance the performance. Initially, this architecture was implemented by connecting the microprocessor and cache inside a ceramic package using wire bonding. This required custom SRAMs and expensive packaging. The implementation evolved to the use of a cartridge form-factor whereby commodity PBSRAMs were connected to the microprocessor by using a printed circuit board.

In the later silicon technology generations, improved Very Large Scale Integration (VLSI) density made it practical to integrate the cache into the same microprocessor chip. This accomplished two major objectives:

- It lowered the cost by eliminating the need for external PBSRAMs and the cartridge.
- It gave higher performance because of a full-speed BSB integrated into the silicon.

This is the current trend for future microprocessors. Consider, for example, the die shown in Figure 10. These are similar die except some have integrated caches and some don't. For a small increase in die size, it is possible to accomplish the two objectives mentioned above.

As silicon features shrink, this mitigates the initial chip size penalty of adding the L2 cache.



(a) (b)

Figure 10: P6 architecture die with and without integrated L2 cache

Future microprocessors may also integrate part of the chipset into the same silicon thereby further reducing the interconnect complexity and costs at the system level.

Driver #2: Connecting the Bus

Although the incorporation of the L2 cache alleviates data traffic on the microprocessor bus, an increase in the bandwidth of the microprocessor bus is still necessary in many applications where I/O bandwidth is important, such as graphics and servers.

To enable high performance, the microprocessor bus speed evolved from an 8 MHz Industry-Standard Architecture (ISA) bus on the original PC-Advanced Technology (AT) to a 100 MHz bus on today's microprocessors. Moreover, there are clear indications that we need to further increase this speed and bandwidth. Aside from the raw speed, additional challenges in high-end systems include the use of multiple processors on a single microprocessor bus. This requires the support of several electrical loads on the same bus thereby necessitating very precise electrical designs to achieve the desired performance. Figure 11 illustrates this configuration.

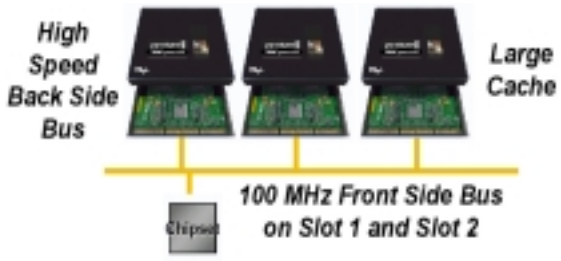


Figure 11: Multiprocessor on a microprocessor bus

To manage these high-performance electrical environments, the interconnections from the silicon through the package to the system board must be designed as an integral unit in order to ensure the desired electrical characteristics. From a technology viewpoint, this requires high-conductivity interconnect traces, low capacitance, and matched impedance for high I/O speed to minimize noise generation. The design of the I/O drivers/receivers on the silicon must also account for package as well as system interconnects. Careful matching of impedance, voltage signal levels, and timing is essential to guarantee performance.

Driver #3: Power Delivery

The third driver is delivering power to the chips. A key element that enabled the advances in silicon technology and the resultant density and performance improvements from generation to generation is the scaling of the supply voltage. While this approach is beneficial to silicon scaling and power dissipation, the challenge of delivering power to the silicon chip via the package is increased. There are two elements to power delivery:

(a) DC (average) supply current

As the supply voltage was scaled, the integration of additional functions and operations at higher clock speeds kept the power dissipation high. As a result, the average supply current increased significantly. This high current was delivered from the power supply on the system motherboard to the chip through the package. As illustrated in Figure 12, the typical supply currents were in the 10 – 20 A range, a range that is expected to increase for future processors. To handle this high current, the package must provide a very low resistance path, in the order of < 1mili-ohm.

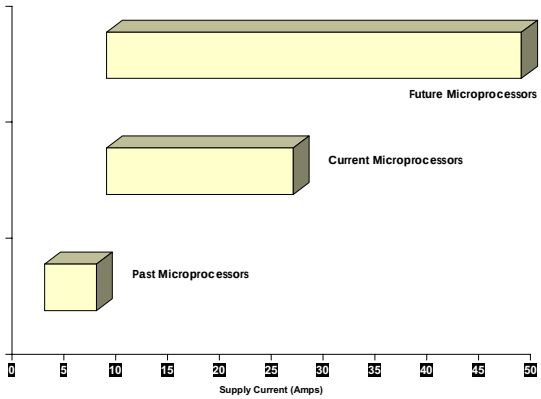


Figure 12: Supply current trends for microprocessors illustrating the effects of voltage scaling and ever-increasing performance

(b) AC (di/dt switching) current

Even more challenging is the management of the switching current. The high clock speed circuits and power conservation design techniques such as clock gating and SLEEP mode result in fast, unpredictable, and large magnitude changes in the supply current. The rate

of change of many Amps per nanosecond of this switching current far exceeds the ability of the power supply and the voltage regulator to respond. If not managed, these current transients manifest themselves as power supply noise that ultimately limits how fast the circuits can operate. This is further compounded by the reduced noise margin in the Complementary Metal-Oxide-Silicon (CMOS) logic circuits that result from power supply voltage scaling.

To mitigate this undesirable noise effect, the package must provide a very low inductance path for the switching current. In addition, charge storage devices, in the form of capacitors native to the silicon chip and augmented by capacitors on the package, are also necessary in some designs.

Driver #4: Dissipating Power

Dissipating high power, and managing high-power density, is another challenge. With the high density of integration and high clock rates, advanced microprocessors dissipate a significant amount of power in a very small physical area. Figure 13 illustrates the problem .

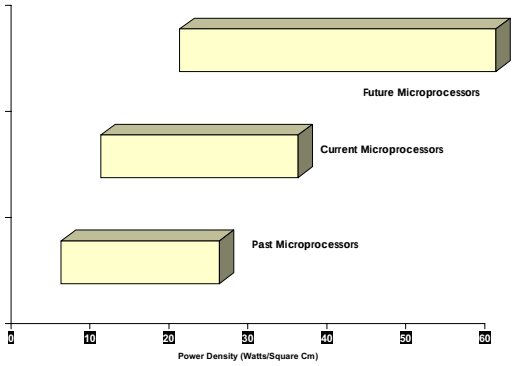


Figure 13: Power density trends of microprocessors

Another factor that exacerbates the thermal management problem is that local areas of the die, depending on where different functions are executed, have higher power densities than the average power density. The challenge to packaging is to ensure that the thermal path from die to the environment is optimized to allow for effective spreading and ducting of heat to the environment. In a broad sense, thermal management involves

- accurately estimating the power dissipation requirements, including power, on-die power distribution, and die temperature expectations,
- managing the thermal performance of all interfaces in the thermal path from die to the environment,
- providing and optimizing the spreading characteristics of all the thermal elements from die to the environment including the package heat-spreader and the heat sink, and finally,
- managing the thermal environment in the chassis by ensuring that the local air temperature is as low as possible to provide a better environment for the microprocessor to dissipate heat.

It is clear that the increasing thermal challenge requires advanced thermal management to ensure chip functionality and reliability.

Driver #5: Silicon Density and Die Shrinks

As silicon technologies advance, the size of the physical feature that can be fabricated gets smaller. Correspondingly, a given amount of circuitry can be built in a smaller die area. Both Intel and the rest of the semiconductor industry employ an aggressive die shrink or die compaction strategy to reduce the silicon area. This approach has two major benefits. First, by reducing the die area, more die can be fabricated on the wafer, hence the cost is lower. Second, a reduction in die area results in higher speed and lower power dissipation for the same speed. This trend is illustrated in Figure 14.

Figure 14: Die shrinks driven by advances in silicon technologies necessitate corresponding improvements in chip-to-package interconnectivity

As the die size shrinks, the number of I/O connections does not change. Furthermore, the number of power supply connections is often increased to support the performance increases brought on by the die shrink. These factors result in a decrease in the bonding pitch for wire bonding or bump pitch for flip-chip. In order to keep pace with this trend, the package geometries and the assembly technologies must also evolve. Today, very fine pitch wire bonding has brought wire bonding down to a pitch of 65 microns. The pitch used on flip-chip arrays is considerably larger, currently in the range of 200 microns, as it utilizes the entire surface of the die to lay out the bumps. Nevertheless, this bump pitch still has to be scaled to keep pace with silicon scaling and die size reduction.

Driver #6: Socketability

Socketability is a business requirement. The reasons for socketability include Original Equipment Manufacturer (OEM) inventory control, the impact of tax and duty, and manufacturing flexibility.

From a technical standpoint, socketability is not desirable. In most cases, it makes the package larger, more expensive, and the performance is lower. Nevertheless, the quest for a low-cost, high-performance socketable package is a strong business-driven requirement.

THE TECHNOLOGY AGENDA

To meet these challenges, the Assembly Technology Development group within Intel has been engaged in defining and creating many new technologies to serve as building blocks as well as integrating the design and analysis environments. The key building blocks are as follows**:**

1. A packaging technology that has high electrically conductive metallurgy that minimizes the IR drop and acts as a high-current conduit to deliver power from the power supply to the chip, such as copper conductors in organic packages.
2. Low-inductance connections from chip to the package and from the package to the socket, such as flip-chip interconnects.
3. Low-capacitance insulator materials, such as organic packages.
4. Advanced thermal-interface materials and a focus on thermal design to manage the high-power density.
5. An integrated analysis, design, and validation environment that enables dynamic trade-offs between chip and package design and layout in the interconnect continuum that includes Computer Aided Design (CAD) tool suites, test vehicles, etc.
6. Predictive models especially in power delivery, power dissipation, and Electromagnetic Interference (EMI).

These building blocks have been in development at Intel for the past several years. Accordingly, a number of significant technology transitions is already underway.

Transitions

Away from Wire Bond and Ceramics

Versatile, ceramic package technology can be expensive. Furthermore, as performance increases, the physical characteristics of ceramic packages may become limiting. Specifically, a ceramic material based on Al_2O_3 has a relatively high dielectric constant ($\varepsilon_r \sim 7$-8). Additionally, because of the high-temperature processing, metallization is limited to refractory metals, such as Molybdenum and Tungsten, which are quite resistive. Wire bond connections have relatively poor electrical characteristics because of their high inductance. As described above, high-resistive, inductive, and capacitive structures are not conducive to high performance. Ceramic substrate suppliers are addressing some of these limitations with advances in their materials.

Use of Organic Packages

A key thrust pursued at Intel, was the transition from ceramics to organic laminate packages. It started in 1996 with the introduction of the Intel® Pentium® processor in the Plastic Pin Grid Array (PPGA) package. Today, all of Intel's microprocessors are in organic laminated packages.

In contrast to ceramic packages, organic laminated packages utilize epoxy resin dielectric materials ($\varepsilon_r \sim 4.2$) and copper conductors. These low dielectric characteristics and copper result in substantial improvements in power distribution and signal transmission characteristics.

The organic package is also indigenous to our latest flip-chip packaging. The attributes of this package provide significant boosts to power distribution and signal routing on the chip. The table below contrasts the physical and electrical characteristics of a typical silicon circuit versus that of a flip-chip OLGA. In short, the routing density is much higher on the silicon, but the electrical characteristics are much better on the organic package. Hence, a judicious use of these capabilities in an interconnect continuum can result in optimal product performance and cost.

Conductor	Pitch (µ)	Material	Thk (µ)	Sheet Rho (mΩ)	Insulator
Si	0.5	Al-Cu	0.5	~85	Oxide
C4-OLGA	70	Cu	17	~0.01	Epoxy

In order to meet the tight pitch demands for chip-area array interconnect (C4 discussed in next section), it was necessary to construct a new organic package. This package uses a laminated core set of layers with high-density "build-up" layer(s). The high-density layers are used to match the pitch on the die. This package is illustrated in Figure 15.

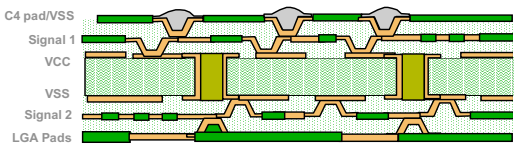


Figure 15: Cross-sectional view of organic flip-chip package illustrating core and high-density build-up layers

Refinements to this technology have been developed to allow alternative package form-factors, as described earlier, based on market segment needs.

Another advantage of organic packaging is that the Coefficient of Thermal Expansion (CTE) of the package is more closely matched to that of the motherboard as compared to a ceramic package. This ensures that the stresses induced in the package-to-motherboard interconnects are significantly lower, resulting in more reliable interconnections even when the interconnect count is large. This is especially true of Ball Grid Array (BGA) connections where interconnect reliability is of significant importance.

C4 Flip-Chip

In order to optimally access the superior electrical characteristics of the organic package, we must also establish a high-density, high-performance method to connect the chip to the package. To that end, a solder-bump-based C4 area array flip-chip capability was developed to replace wire bonding.

In contrast to traditional wire bonding, C4 utilizes an area array method for interconnection. The C4 bumps can be placed anywhere on the die, even over active circuitry. This enables the placing of many more bumps as virtual vias (through the thick electrical connections) connecting the metallization on the chip to the metallization on the package. In fact, the metallization on the package can be visualized as metal layers augmenting the metal layers on the chip. The primary benefit of this approach is in power/ground distribution and clock routing. The C4 connections, in combination with the electrical characteristics of the copper-based organic packages, result in a superior electrical environment where maximum performance can be realized. As an example of this implementation, consider the same die, P6 architecture, in both the wire-bonded and C4 versions as shown in Figure 16. The substantial increase in power/ground connections ensures maximum performance. Additionally, a native C4 die design eliminates the need for bond pads, which results in a small die, ~0.012 inches smaller per side.

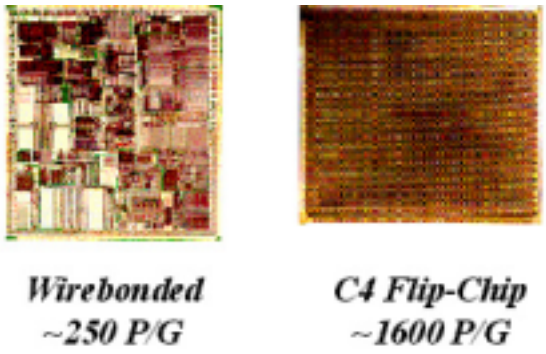


Figure 16: P6 architecture microprocessor implemented as wire-bonded die and in C4 Flip-Chip

In this paper, we discuss packaging and technology building blocks as a key concept to cost effectively meet a wide range of both form-factor (surface-mount, high-density pinned, low-density pinned, etc.) and performance needs for packaging by a judicious combination of these building blocks.

Figure 17 further illustrates this concept. A single bumped die is mounted on either a surface mount (OLGA) or a pinned substrate. The OLGA substrate can subsequently be surface mounted directly to a board, mounted to a pinned substrate for socketing, or mounted to a cartridge (to be combined with other chips).

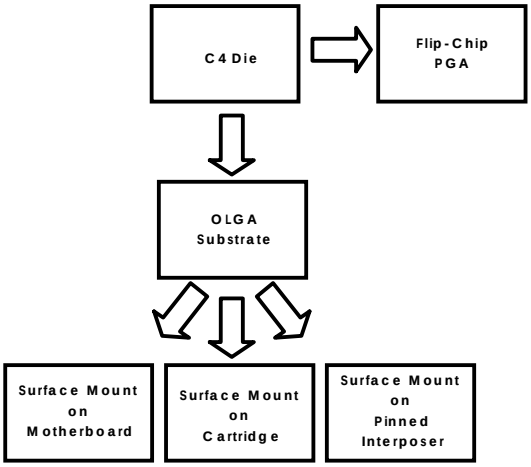


Figure 17 : Building-block technologies

FUTURE CHALLENGES

As microprocessors continue to improve in performance, technical challenges in packaging will also increase. A comprehensive view of these challenges can be found in the 1999 International Technology Roadmap for Semiconductors (ITRS) [1]. This roadmap discusses the need for improved materials and assembly processes as well as a need to have integrated simulation tools and

methods to assess the reliability of the integrated die-package-motherboard system. Since the design environment and the assembly processes and reliability aspects of packaging fall outside the scope of this paper, we limit this discussion to the technical challenges in packaging as they impact microprocessor performance. Technically, the challenges fall into three broad categories: power delivery, power removal, and also the provision of viable, i.e., appropriately scaled with optimal performance characteristics, interconnection strategies between silicon and board.

Power delivery challenges are highlighted in Figure 12. To move forward, the focus will have to be on continuing to understand and optimize the electrical path from power supply to the die. With increasing demand for performance, the general separation of market segment requirements and constraints, and the shortening of the time-to-market, it is expected that the power delivery solutions will continue to be challenging.

Power removal, i.e., thermal management of the processor, is another increasingly challenging aspect of packaging. As shown in Figure 13, the average power density of processors is expected to increase. The problem will be exacerbated by the need to manage local power densities on die. The development of cost-effective and technically viable thermal management solutions that maintain die temperature at acceptable levels will be key to ensuring future success. This can be accomplished through development and deployment of effective spreader solutions and thermal interface materials. Controlled assembly processes to manage the thermal interfaces are also a key to successful design. Finally understanding and managing the die power, power distribution, and the thermal environment in the chassis are important.

Silicon technology in the future is expected to scale aggressively, which will require intelligent space transformation methods from packaging. Ensuring that the interconnects have refined electrical characteristics so that packaging provides the appropriate space transformation while enabling the required electrical performance will be essential to future development.

SUMMARY

In this paper, we discuss the evolution of microprocessor packaging from a simple protective enclosure to a more technically complex and challenging platform that enables optimal microprocessor performance. The general strategy adopted within Intel to address continuing challenges is to develop building blocks that can be effectively combined to meet the needs of current and future microprocessor packaging. The remaining papers in this issue of the *Intel Technology Journal* discuss different aspects of these challenges in greater detail.

The 2nd paper discusses the FC-PGA package, i.e., flip-chip technology on organic pin grid array substrates. The paper looks at the motivations that led to the development of this package technology, its characteristic features and capabilities, and some of the issues that were successfully addressed during the development and deployment of this technology into high-volume manufacturing.

The 3rd paper discusses the technical complexity of interconnect design to achieve optimal electrical performance. This paper also discusses the design analysis and synthesis techniques used to ensure optimal electrical design.

The 4th paper discusses challenges in power removal. Thermal solutions that are optimized for cost and performance and tailored to meet different market segment needs are a key enabler to successful microprocessor deployment. This paper discusses some of the considerations that must be taken into account to ensure successful thermal design.

Underfill processes and underfill material development are a major component of flip-chip packaging processes. The 5th paper discusses a novel method of accomplishing this objective.

Ensuring that packaging continues to meet high standards of reliability is a key to success. The 6th paper discusses how our assessment of reliability has evolved during the past few years. Intel has moved from a stress-based certification strategy to a more fundamental mechanism-based methodology that allows for a better linkage between stress testing and the end-user environment.

Finally, the 7th paper talks to the practical problem of managing the thermal environment during microprocessor testing. The goal of testing is to effectively assess performance and reliability without introducing artifacts due to the testing process. This paper examines how this goal can be accomplished and looks at some of the unique issues that should be addressed.

CONCLUSIONS

High-performance and cost-effective microprocessor technologies require a holistic approach that comprehends the interconnect continuum including the silicon, the package, and the system. By properly exploiting the attributes of these regimes, optimal performance and cost can be realized. This review of the evolution of packaging reinforces the view that it will be a technically challenging and rewarding area of focus.

ACKNOWLEDGEMENTS

The authors thank Bill Siu for providing the essential framework of this paper. We also acknowledge the support of Scott Shirley, Susan Coartney, and Keith Carnes in procuring some of the graphics for this paper.

REFERENCES

[1]"International Technology Roadmap for Semiconductors," 1999 Edition, *Semiconductor Industry Association*.

AUTHORS' BIOGRAPHIES

Ravi Mahajan is currently the manager of a Thermal and Mechanical Analysis group in Assembly Technology Development. His group provides design and analysis support in the development of advanced packaging for Intel processors and chipsets. He also leads a group of thermal experts from different organizations within Intel. This group is chartered with developing and maintaining the thermal roadmap for Intel microprocessors. Dr Mahajan received his B.S. degree in Mechanical Engineering from the University of Bombay (1985), his M.S. degree in Mechanical Engineering from the University of Houston (1987), and his Ph.D. degree in Mechanical Engineering specializing in fracture mechanics from Lehigh University (1992). Dr. Mahajan holds several patents in packaging, has edited two conference proceedings for the Society of Experimental Mechanics, and is currently one of the editors of an Intel Internal Technology Journal. His e-mail is ravi.v.mahajan@intel.com.

Ken Brown is manager for non-CPU Packaging Development in the Assembly Technology Development group at Intel. He has worked in the semiconductor industry for 19 years, and currently serves as Intel's representative to the SIA International Technology Roadmap for Semiconductors (ITRS) packaging committee. His first 17 years were with Digital Equipment Corporation, where his last position was as Manager of the Package Design & Development organization. He moved to Chandler and ATD about two years ago following Intel's acquisition of Digital Semiconductor. He received a B.S.M.E. degree (1981) from the University of Massachusetts, an MBA from Northeastern University (1985), holds two patents in packaging, has presented at several packaging conferences, and authored a chapter on packaging for a graduate text book. His e-mail is ken.m.atd.brown@intel.com.

Vasu Atluri is currently a Silicon Integration Manager in Assembly Technology Development. His group is responsible for ensuring that all technical issues are properly managed to enable a successful interface between silicon and assembly technology development groups. He is also responsible for the design of test chips that help validate performance and reliability of silicon after it has been packaged. He also helps in defining design rules for product designs.

Dr Atluri received a B.S. in Chemical Engineering from Osmania University (1982), anM.S. in Materials Science and Engineering from University of Arizona (1987), an M.S. in Chemical/Metallurgical Engineering (1988), and a Ph.D in Materials Science and Engineering from University of Arizona (1998) with a specialization in silicon fabrication. Dr. Atluri has more than twenty technical publications in conference proceedings and journals. He has presented at several technical conferences. He has filed for couple of patents in silicon processing. Dr. Atluri is currently one of the editors of Intel's Assembly and Test Technology Journal, which is Intel Corporation's Internal Technology Journal. Dr. Atluri is an active member of IEEE and is currently serving as Vice-Chair for IEEE Phoenix Section. Dr. Atluri maintains active participation in a wide array of academic forums associated with electronic packaging. His e-mail is vasudeva.atluri@intel.com.

Simultaneous Chip-Join and Underfill Assembly Technology for Flip-Chip Packaging

Tom Dory, Assembly Technology Development, Intel Corp.
Kenji Takahashi, Japan Package Technology Development, Intel Corp.
Tomomi Kume, Japan Package Technology Development, Intel Corp.
Jiro Kubota, Japan Package Technology Development, Intel Corp.
Seiichiro Seki, Japan Package Technology Development, Intel Corp.
Takaharu Fujiyama, Japan Package Technology Development, Intel Corp.

Index words: C4, underfill, packaging, flip chip

ABSTRACT

Controlled Collapse Chip Connection (C4) flip-chip packaging is the current state of the art assembly technique. The C4 Organic Land Grid Array (OLGA) technology process to attach a silicon die to a package substrate involves seven process steps.

In the current plan of record (POR) C4 OLGA flip-chip assembly process, solder wetability is achieved by removing Pb/Sn oxide films from the bump material by means of a flux and deflux process. An epoxy material is placed between the die and package. This underfill (UF) process has a long epoxy application step and curing time.

We present a three-step chip-join assembly process using a no-flow-type underfill material combined with non-lead-containing bump metal on the chip side. This process has fewer process steps than the POR assembly process, and the underfill cure time is also reduced.

The process described in this paper shortens the assembly process by eliminating fluxing, defluxing, and furnace reflow steps. We achieved more than a 90% assembly yield as a result of process parameter and material property optimization.

INTRODUCTION

Intel successfully introduced C4 OLGA technology into the Personal Computer (PC) market in April 1998. The technology has many advantages over wire bonding and Tape Automated Bonding (TAB): it has a higher Input/Output (I/O) number, shorter interconnects, and the silicon die self-aligns to the package.

The current C4 OLGA plan of record (POR) process to attach a silicon die to a package substrate and apply the underfill material requires seven processing steps. An epoxy-type material is required in C4 packaging to prevent solder bumps on the silicon die from moving and electrically shorting during the life of the part. The underfill also mediates the thermal miss-match between the silicon die and the organic polymer package. The underfill material is applied in a liquid form to flow between the silicon C4 bumps and the package substrate. A high-temperature oven cure is required to set the polymer.

Several different underfill formulations and bump metals were examined. We found a correlation between the size of particles in the underfill material and silicon bump to package substrate joint reliability. Different underfill resins were tested. Resins causing a higher viscosity showed better reliability after a temperature cycle reliability test.

We examined the potential package assembly cost savings using our three-step process. By eliminating process steps and shortening the cure time, a substantial cost savings can be realized.

This paper describes an assembly technology that targets future-generation C4 flip-chip packaging. We report a method to combine silicon chip join and epoxy underfill processes in one step with a shorter underfill cure time. The process has fewer steps than the POR assembly process. The POR assembly is shown in Figure 1 with our new assembly process flow shown in Figure 2.

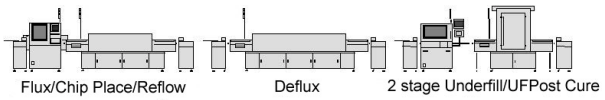


Figure 1: Current (P856) chip join assembly process

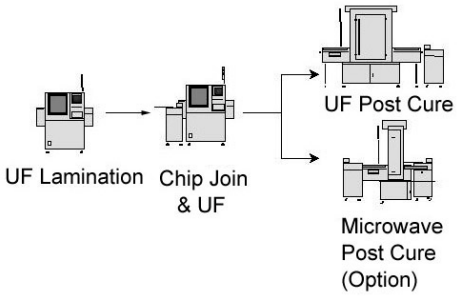


Figure 2: New one-step chip join assembly process

MATERIALS AND PROCESS

Materials

Die

Test die were used with different bump metal material. To obtain good adhesion between the bump material and die, a Ti/Au under bump metallurgy was used.

Substrate

The test substrate was an OLGA material with Sn/3.5Ag eutectic pre-solder paste.

UF Material

The no-flow-type underfill material provided by Nitto Denko was used in the testing. The UF properties of the first formulation are listed in Table 1[1].

Table 1: No-flow-type UF property

thickness	80μm
resin	non-conductive
CTE (below Tg)	41ppm/K
CTE (above Tg)	113ppm/K
gel time at 175 °C	23 seconds
Tg	117 °C
film size	12 x 12mm

Process

The new process has three steps: 1) UF lamination; 2) chip placement and UF cure, and solder reflow with a flip-chip bonder; and 3) post cure of UF. Figure 3 shows the chip-join assembly step. First, UF was manually laminated onto the substrate with a roller on a hot plate heated at 90°C. Second, chip join was performed with a semi-automatic bonding machine. The heat stage was set at 135°C. After chip placement, the temperature was ramped up to 192°C by the use of a head tool to cure the UF for 60 seconds. The temperature was increased to 230°C for solder reflow. Post-cure cool down was conducted at 150°C/1hr using a conventional oven.

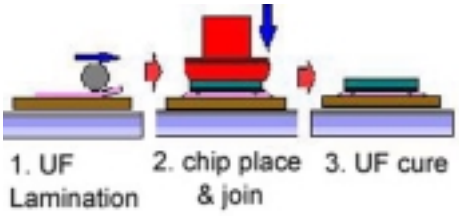


Figure 3: One step chip-join assembly process

Microwave Oven Reflow and Cure Process

To further reduce process time, we used a microwave oven to cure the UF epoxy. This replaced the conventional oven and flip-chip bonder reflow process. Using a microwave to cure the epoxy took less time than a conventional oven and a flip-chip bonder reflow process as shown in Figure 4. Therefore, assembly time will be reduced from 62 to 8 minutes per unit when the microwave cure is used in this new chip-join assembly process.

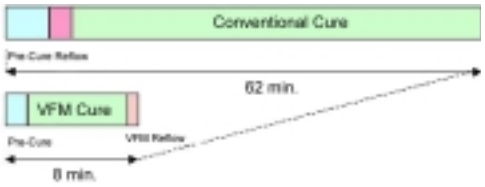


Figure 4: Microwave reflow and cure process for one-step chip-join assembly process

All samples using the microwave cure process passed the initial electrical test at the interconnection side of the solder joint. The sample size was 955 pieces in five substrate lots. However, we found that 9.2% of the failures on the substrate side of the solder joint were caused by an electrical open due to substrate cracking. Additional process characterization is required for microwave epoxy curing.

RESULTS

During development of this new assembly process, we found a joint problem after assembly. Assembly yield loss was due to an electrical open found in the peripheral area of the die.

Solder Flow-Out

Figure 5 and Figure 6 show the cause of the electrical open: solder flow-out from the interconnection.

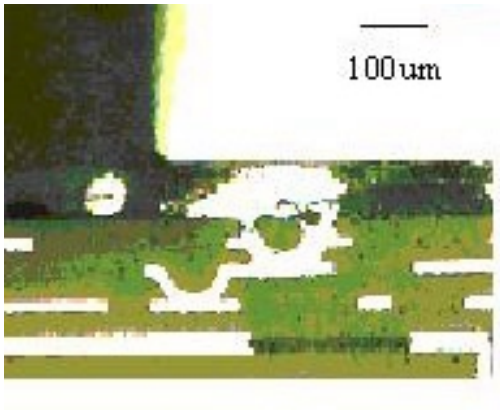


Figure 5: Solder flow-out from interconnection

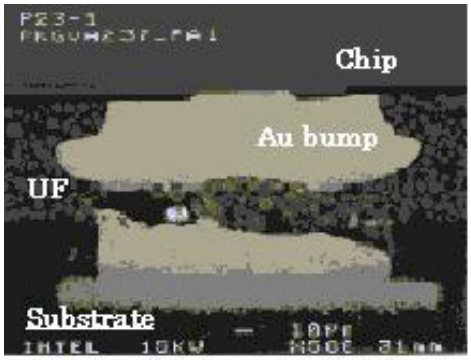


Figure 6: Solder flow-out from interconnection

Root Cause Analysis of Solder Flow-Out

Figure 7 shows the X-ray image of the solder flow-out from the interconnection before and after reflow. From the photo, it was found that solder flow-out happened during reflow. As a first step to eliminate this defect, the bonding force during reflow was reduced from 30kgf to 2kgf. An X-ray showed that solder flow-out was reduced by reducing the bonding, but not significantly.

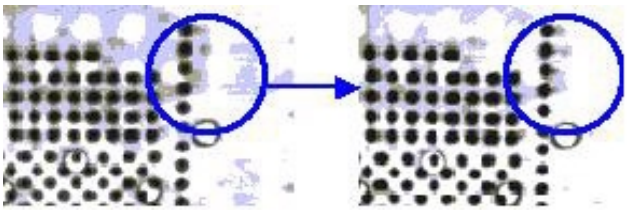


Figure 7: Solder flow after reflow

To determine the root cause of solder flow-out, the condition around bumps before reflow was examined. Figure 8 shows an X-ray image of the solder bumps on the silicon die.

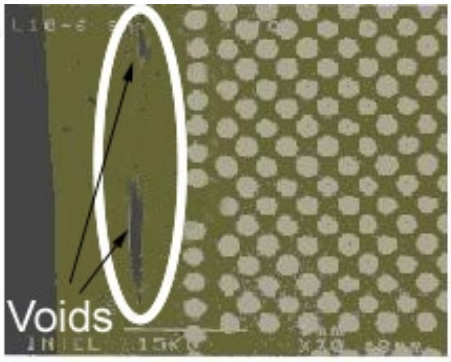


Figure 8: Underfill voids before reflow

The root cause of solder flow-out was void generation at the side of the bumps during die placement/cure, which was most likely caused by moisture absorption in the substrate. An experiment was conducted to compare the effect of substrate prebake against solder flow-out. Figures 9 and 10 show the X-ray views of the silicon bumps with and without substrate prebake. A significant reduction in voids was achieved with substrate bakeout; however, the voids were not completely eliminated.

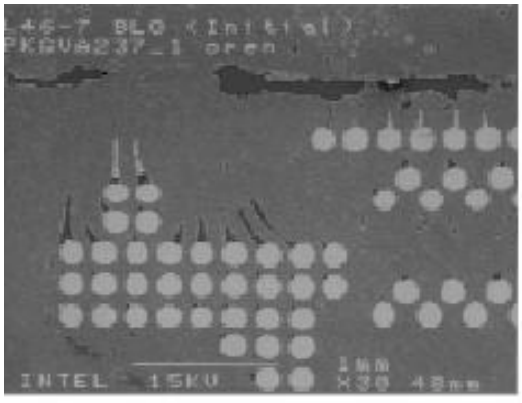


Figure 9: X-ray image of initial sample without prebake showing voids (dark areas) in the underfill material

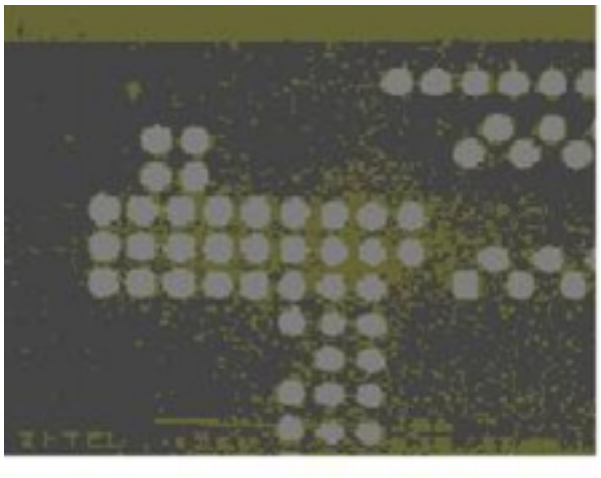


Figure 10: X-ray image of initial sample with prebake showing solder bumps

Rev 1 Baseline Assembly

We collected data for a Rev 1 baseline. Table 2 shows process conditions.

Table 2: Rev 1 baseline assembly process profile

Process Step	Condition
UF lamination	90 °C
UF cure/chip join	135 °C ramped to 192 °C for pre-cure, ramped to 228 °C for chip join
UF post cure	150 °C
total time	60 minutes

DISCUSSION

The assembly result showed 13/3000 pieces failed at the initial testing before reliability testing. Figure 11 lists the cause of each failure.

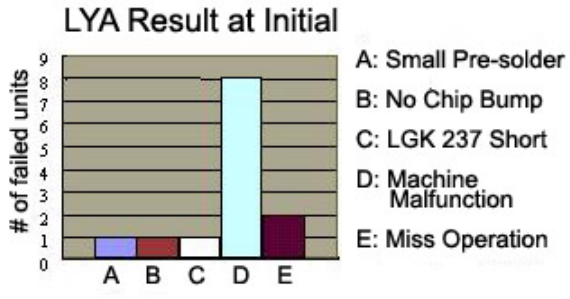


Figure 11: Failure analysis at the initial reliability test

Reliability Evaluation

Rev 1 reliability test results are summarized in Table 3. There were no electrical short failures after 100 hours of performing the Highly Accelerated Stress Test (HAST). We did observe one unit failed because of an electrical open. The open was due to poor interconnection contact that was caused by insufficient presolder volume. We believe the reason for the excellent HAST performance is the elimination of Pb in the interconnect materials.

Table 3: Rev 1 baseline reliability yield

Test	# Units Failed	Yield (%)
initial yield before reliability testing	13	95.7
T/S-B 500 cycles	5	91.7
T/S-B 1000 cycles	49	18.3
pre-condition T/C-B 500 cycles	55	0
HAST 100 hrs.	1	98.2
HAST 200 hrs.	3	94.5
steam 168 hrs.	13	82.9
steam 336 hrs.	63	17.1

The key issue of this first reliability test is the temperature cycle-B [2], T/C-B, 1000 cycle test, in which 91.7% of the samples failed due to electrical high resistance. Figure 12 shows the interface crack of the connection after T/S-B 1000 cycles.

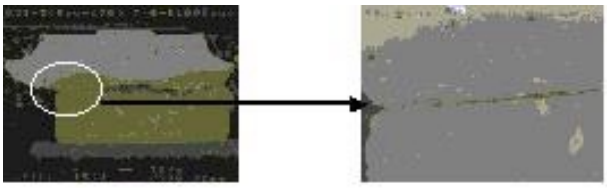


Figure 12: Interconnection crack after T/S-B 1000 cycles

Thermal Stress Issues and Improvement

In the Rev 1 baseline testing, almost all samples failed the T/C-B thermal cycle test. To improve reliability, we first modified the UF material. In the first UF modification, we compared crystal resin A with multifunctional resin B. Using a B-type resin, the UF had a higher yield due to less thermal stress between the two resins. In addition, we also compared viscosity values. Table 4 shows the electrical yield between resin types A and B.

Table 4: Reliability result for changing UF properties

UF type	A	B	Rev 1 UF
viscosity	220	300	370
T/C-B 1000 cycles	100	100	18.3
pre-condition T/C-B 500 cycles	36.8	70.0	0

Underfill A contains a crystalline resin. Underfill B and the Rev 1 underfill contain a multifunctional-type resin.

Resins A and B UF passed the T/C-B 1000 cycle test, but failed the precondition TC-B test. Figure 13 shows the relation between viscosity and percent failures.

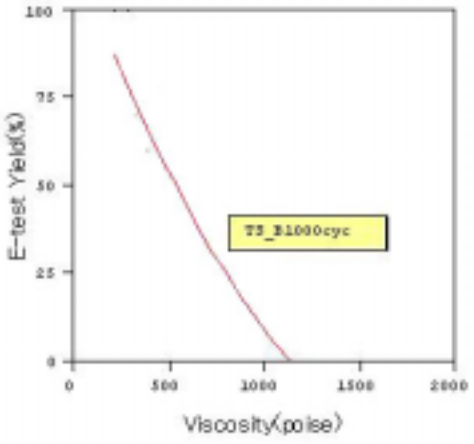


Figure 13: Yields before and after T/C-B vs. underfill viscosity

In the second UF modification, we changed filler size from 5 to 1µm. Filler residue was observed in the interface part between the die bump metal and substrate solder. Figure 14 shows the one- and five-micron filler residue at the initial condition. The 1µm filler residue in the metal interface is lower in quantity than the 5µm filler in spite of the same filler quantity being in the UF.

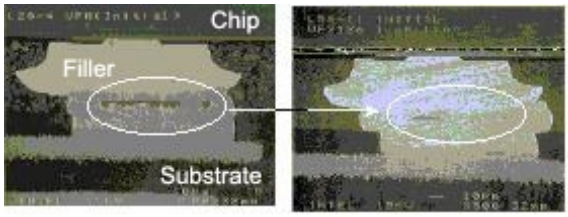


Figure 14: Filler residue at interface by filler size

We collected data on the 1µm filler and low viscosity UF for two UF materials, crystal (C) and multifunctional (D), resin. Table 5 shows that the units using 1µm filler UF showed an improved yield compared to previous baseline using 5mm filler UF.

Table 5: Data collection result of 1mm filler UF

UF type	C	D	Rev 1 UF
T/C-B 1000 cycles	100	100	18.3
pre-condition T/C-B 1000 cycles	69.4	83.3	0 (500 cycles)
interconnection side yield	100	100	

All units passed after the PCTC-B 1000 cycle test on the interconnection side. We observed a failure due to an electrical failure within the substrate side, not the interconnection side. There was no significant difference between crystal and multifunctional resin in this experiment. However, delamination between chip and UF occurred in both resins C and D more than it did in the Rev 1 baseline experiment. The 1µm filler UF properties were checked in terms of delamination. We assumed that the rubber content within the UF material changed the viscosity and caused delamination in resins C and D as shown in Table 6.

Table 6: Relation between viscosity and void and delamination percent at two viscosities

UF type	C		D		A	Rev 1
UF lot #	1	2	1	2		
viscosity 1	310	410	820	1010	1070	
viscosity 2					220	370
rubber index	75	75	100	100	100	100
void % (initial condition)	0.14	0.09	0.09	0.09	0.13	0.01
void % (after Steam 336)	13.67	8.61	6.68	6.07	0.32	0.13

Viscosity 1: 175°C, 10kgf/cm^2, Viscosity 2: 175°C, 110kgf/cm^2.

We observed a greater percentage of voids than in the Rev 1 baseline at the initial solder-join conditions. This not only caused delamination to occur because of poor adhesion after the steam reliability test, but it also caused poor UF property at the initial prereliability test.

COST DISCUSSION

Figure 15 shows the cost model estimation by Assembly Capital Equipment Development (ACED). It compares

the current POR assembly process with our three-step, no-flow process using standard cure ovens and a microwave cure oven in High Volume Manufacturing (HVM) run rates.

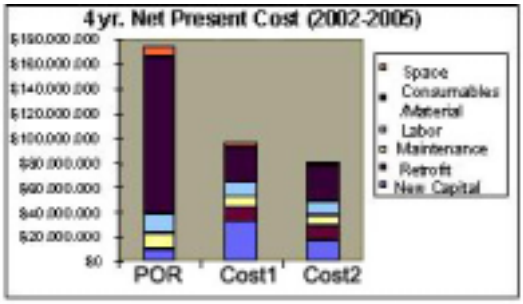


Figure 15: Assembly cost model

The columns are Cost 1: no-flow type underfill with microwave reflow and cure, and Cost 2: no flow type underfill with BTU reflow and conventional oven cure.

Both processes using no-flow type UF reduced the assembly cost when compared to the current OLGA process. We project a reduction in assembly cost over four years of $76 and $92 million when we model Cost 1 and Cost 2, respectively. The main cost savings is expected to come from material cost reduction by using no-flow UF. As a result of this process evaluation, to establish short assembly process steps, we have adjusted the assembly condition and UF material for pathfinding activities. Assembly and thermal stress tests were improved but, the delamination issue, after steam, still remains when 1μm filler UF is used. Both thermal fatigue and delamination are concerns for viscosity quantity in UF. Underfill viscosity is one important factor for thermal fatigue and delamination for this assembly using a no-flow-type UF.

When a low viscosity UF is used, the void and delamination present increased. Void at initial joint formation depends on delamination percent after the steam test. To obtain a low void and delamination percentage, we have to raise the rubber content. However, this may cause thermal fatigue to be increased by high viscosity. We should carefully set the viscosity value in new underfill materials. For future work, we need to evaluate the reliability of UF adjusted viscosity and focus more on void and delamination percentages.

CONCLUSION

A new C4 interconnection/fluxless assembly technology was proposed. The results were promising in that they demonstrated the feasibility of a one-step chip-join process that uses non-oxide forming bumps, Sn/Ag presolder, and no-flow type underfill materials. This process will reduce the number of assembly steps as well as the cost. The assembly yield was more than 90% for this pathfinding activity, and it passed an electrical test after 200 hours of HAST and 1000 cycles of PC T/C-B. We improved the assembly and reliability performance by process and material modifications as discussed.

As the C4 pad pitch shrinks and the gap between die and substrate decreases, the primary concerns are Pb electro-migration and underfilling capability for flow-type underfill materials. Using Pb for finer C4 pitch application will increase the risk of leakage by migration. There also is an effort underway to remove lead from solder in the semiconductor industry. Underfill material is dispensed to the bump gaps by capillary action. With a smaller bump gap, longer process times will be seen and may cause an increase in voids. Large die used for microprocessors present a challenge for void-free capillary underfill flow using the current POR method.

No-flow underfill material may be usable for lower C4 gap, smaller C4 pitch, and large die. An added benefit is the elimination of lead in the solder material. With a nonoxidized bump surface and substrate presolder, a good metallurgical and electrical connection without flux can be made.

ACKNOWLEDGMENTS

The authors acknowledge the support and apt advice of Yukihiro Murakami, Masahiro Kuroda, Koichi Okudaira, Kinya Ichikawa, Kenzo Ishida for JPTD, and Nitto Denko for UF material makers.

REFERENCES

[1] Satoshi Ito, *et al.*, "A Novel Flip-Chip Technology Using Non-Conductive Resin Sheet," Proceedings 48th Electronics Components and Technology Conference, pp. 1047-1051, 1998.

[2] Temperature cycle-B conditions taken from Mil Std. Spec 883E, method 1010.7. Temperatures are cycled between -55 °C and 125 °C.

AUTHORS' BIOGRAPHIES

Tom Dory is the package design integration engineer responsible for the next IA64 microprocessor, Mckinley, and the chipset supporting Mckinley. He earned his Ph.D. in Chemistry at the University of Oklahoma. Since joining Intel twelve years ago, he has had many responsibilities including fabrication equipment development, assembly process manufacturing, and currently new technology package design. He is also the assembly process section editor of the *Intel Assembly and*

Test Technology Journal, an internal publication. His e-mail address is thomas.s.dory@intel.com.

Takaharu Fujiyama is a manager of assembly architecture development #2 in the Japan Package Technology Development Group at Intel K.K. He has an M.S. degree in Electrical Engineering from Kanazawa University. He joined Intel in 1994 and started as a process module engineer of C4 flip-chip assembly technology. He was also involved in the development of the Organic Land Grid Array substrate. The C4 flip-chip on OLGA packaging technology has been used since 1998 for CPUs. His current focus is new assembly architecture development for future-generation CPUs. His e-mail is fujiyamat@intel.co.jp.

Jiro Kubota is a Package Development Engineer in the 1st Assembly Architecture Development group of Japan Package Technology Development at Intel K.K. He has an M.S. degree in Applied Physical Engineering from the University of Miyazaki. Since graduation, he has worked for Intel K.K. and has focused on the development of C4 packaging. He was one of the initial members of the C4 package development working group on JPTD. His e-mail is kubotaj@intel.co.jp.

Tomomi Kume graduated from Tsukuba University in 1997 with a M.S. degree in Science. She began work at Intel in 1997 and then joined the Japan Package Technology Development group as a packaging development engineer. She has focused on the C4 assembly method with flip-chip bonder and film type UF, film type UF material selection, and fluxless chip join. She is currently a mechanical engineer working on the Japan Mobile Support Operations. Her e-mail is kumet@intel.co.jp.

Seiichiro Seki joined Intel in 1992 as an Internal Purchasing Operation engineer. Then he moved to the TMG-J LAB Department in 1994. He has focused on failure analysis for C4 assembly technology with the Japan Package Technology Development group. His e-mail is sekise@intel.co.jp.

Kenji Takahashi is the Assembly Architecture Development #1 manager in Japan Package Technology Development in the Technology and Manufacturing Group, Japan at Intel K.K. He joined Intel in 1996 and has focused on developing flip-chip assembly architecture for future CPU packaging. He holds a B.S. degree in Applied Chemistry from Hokkaido University in Japan. His e-mail is takahashike@intel.co.jp.

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.

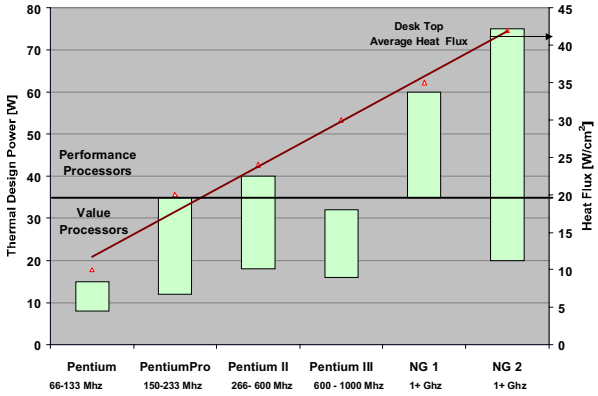
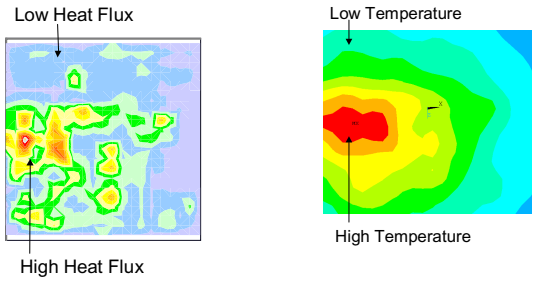


Figure 1: Power and heat-flux trends in the performance and value desktop processors

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.

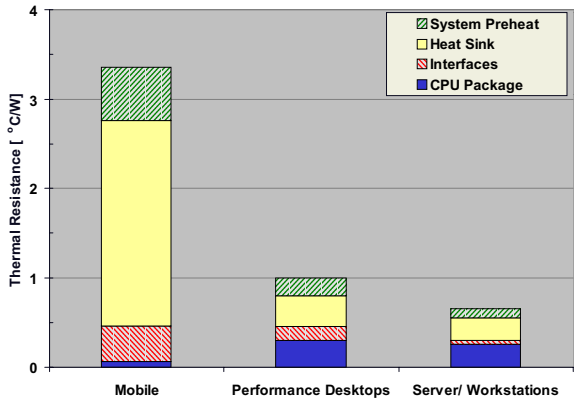


Figure 3: Typical thermal-resistance partitioning (across market segments)

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.

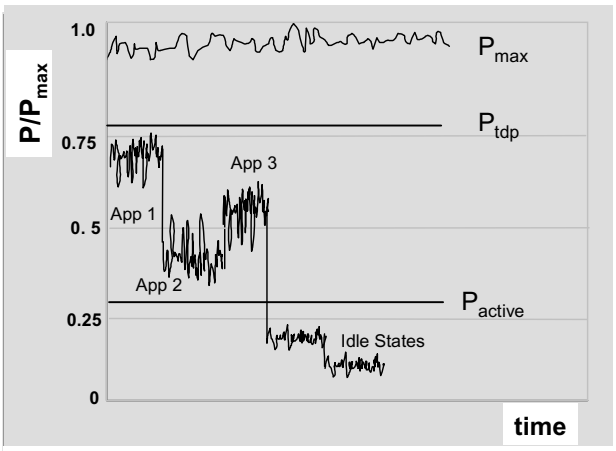


Figure 4: Example of microprocessor power consumption profiles

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.

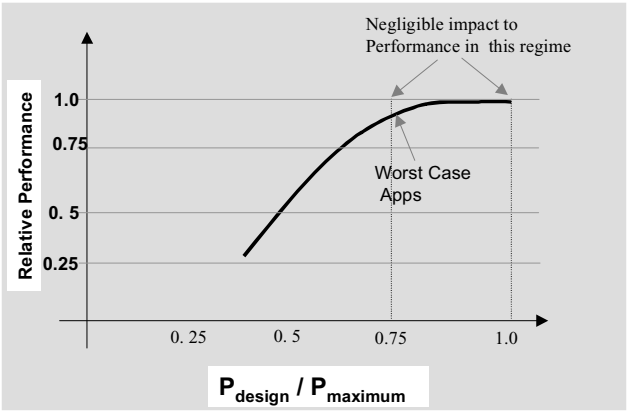


Figure 5: Illustration of impact of thermal design power on processor performance

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.

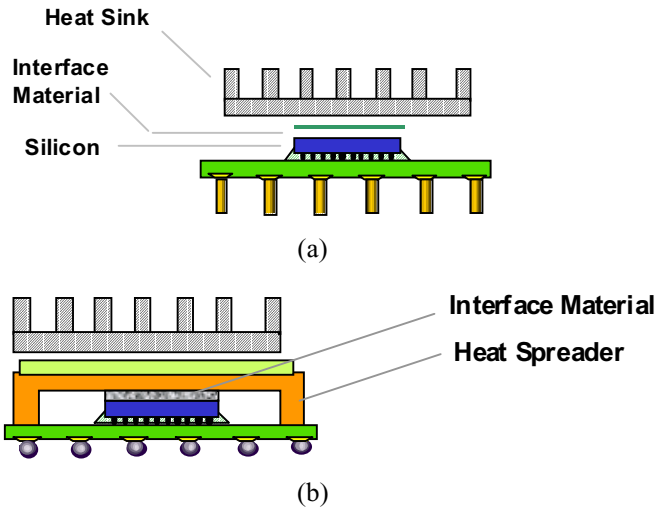


Figure 6: Schematics of FC-xGA1 and FC-xGA2 packaging architectures

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.

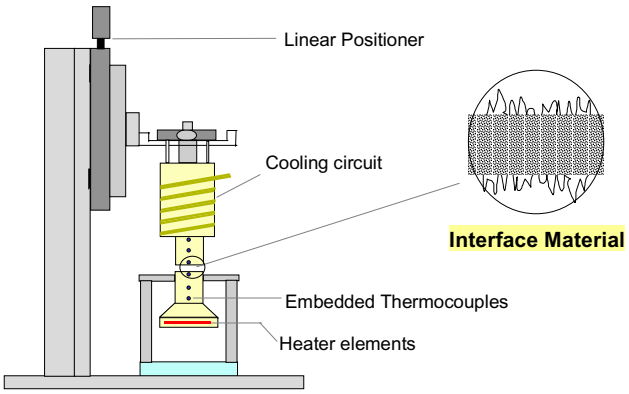


Figure 7: Schematic of set-up used to screen interface materials

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.

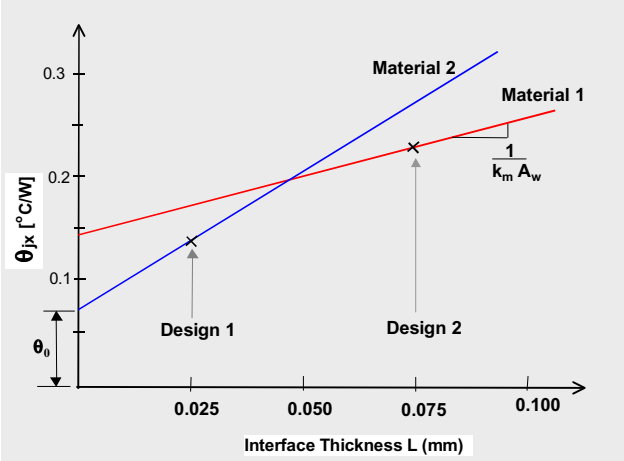


Figure 8: Schematic plot of dependence of thermal resistance to thickness and bulk material conductivity

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 [$\theta o << L/(k_m A_w)$].

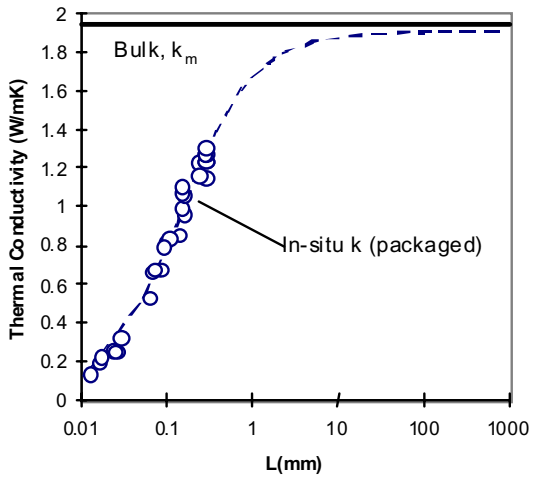


Figure 9: Variation of in-situ thermal conductivity of material (in a package) with thickness

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.

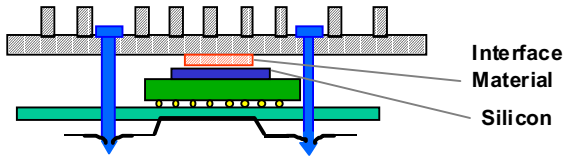
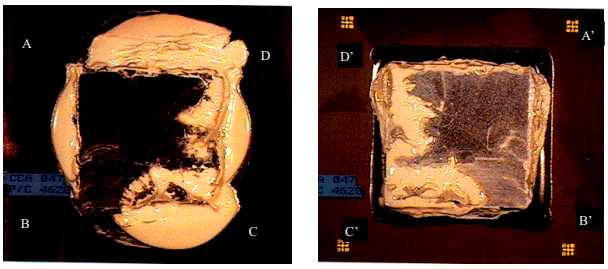


Figure 10: Schematic of the set-up used in thermal grease and phase-change film evaluations

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.

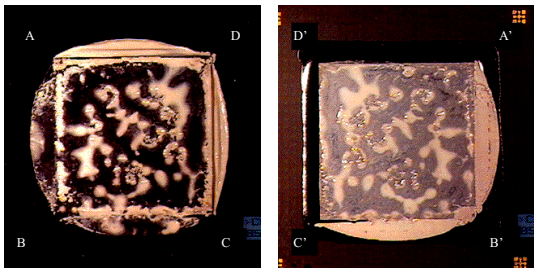
Power Cycling: This is a loss of material due to a phenomenon called "pump-out" (shown in Figure 11). Under cyclic loading, extensive thermo-mechanical stresses exerted at the interface because of the relative motion (flexure) between the die and the base of a heat-sink lead to loss of grease material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

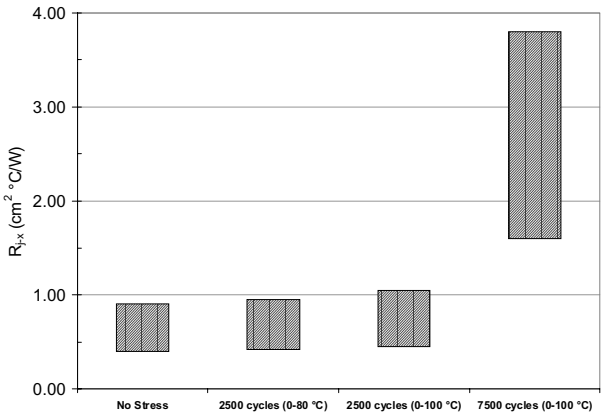


Figure 13: Impact of environmental stress conditions on thermal grease performance

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.

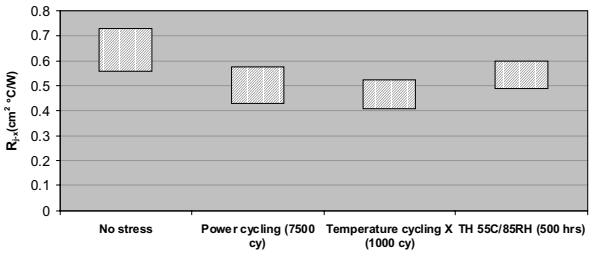


Figure 14: Performance of a PCF material through reliability stresses

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.

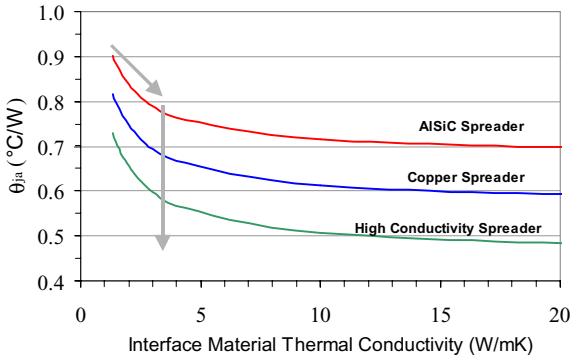


Figure 15: Dependency of thermal resistance with heat-spreader material and thickness

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.

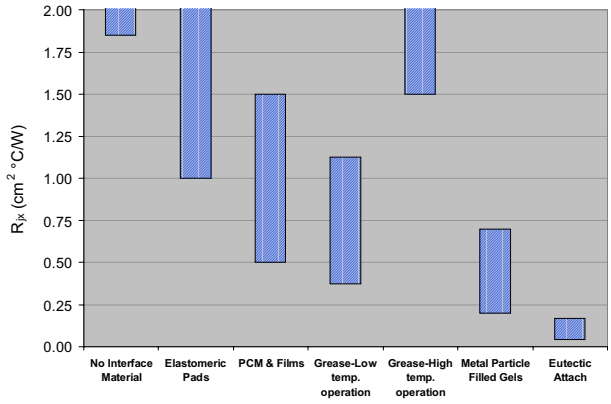


Figure 16: Thermal resistance of typical thermal-interface materials

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.

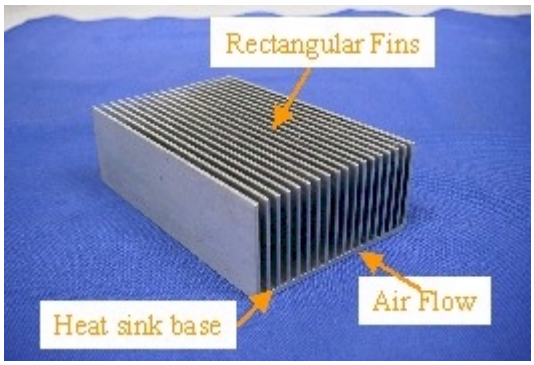


Figure 17: Rectangular fin heat sink

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.

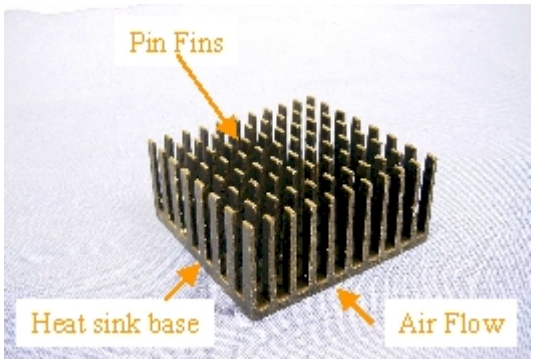


Figure 18: Pin fin heat sink

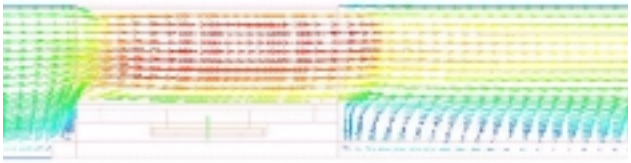


Figure 19: Velocity vectors around a Intel® Pentium® processor heat sink

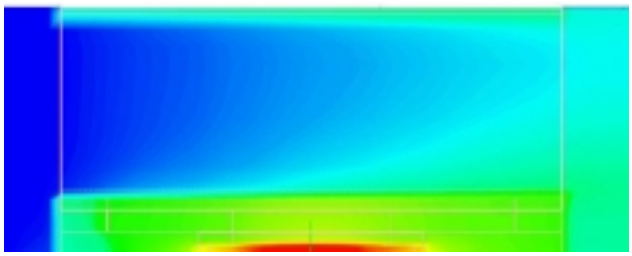


Figure 20: Temperature contours on a Intel® Pentium® processor heat sink

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.

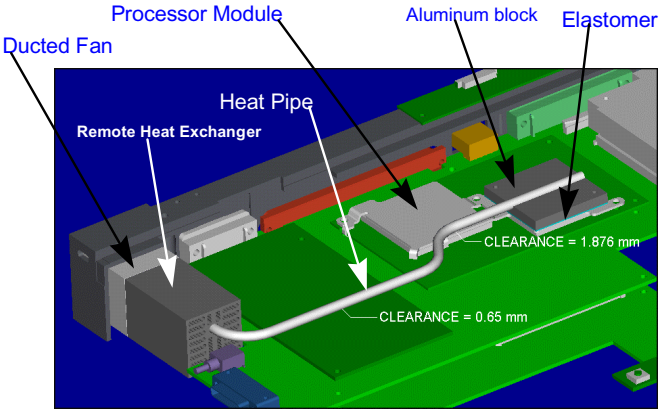


Figure 21: Remote heat exchanger concept for mobile Intel® Pentium® II processor module

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.

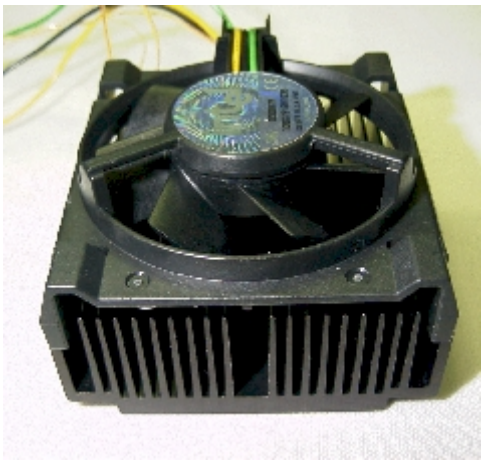


Figure 22: An active heat sink

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.

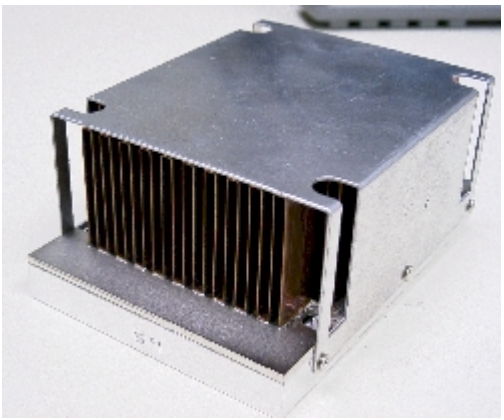


Figure 23: Folded-fin heat sink with a shroud

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.

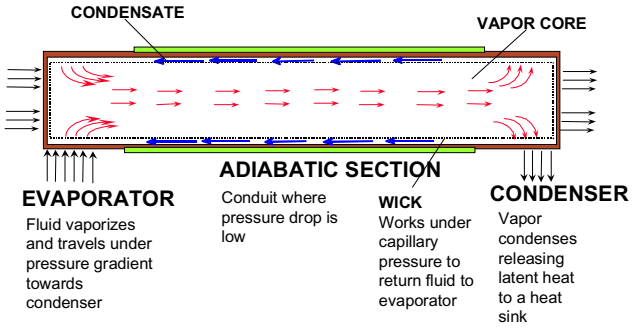


Figure 24: Schematic of a vapor chamber or heat pipe

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.

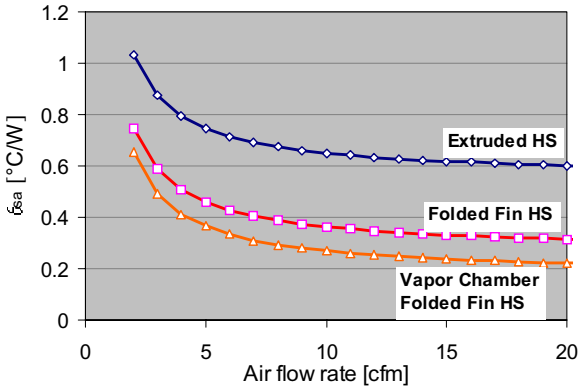


Figure 25: Thermal performance of desktop and server heat-sink designs

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.

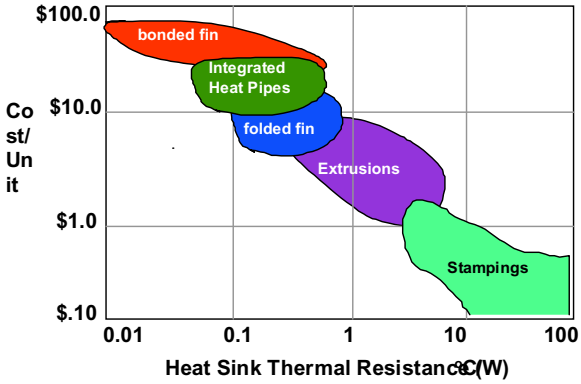


Figure 26: Heat-sink cost vs. performance

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.

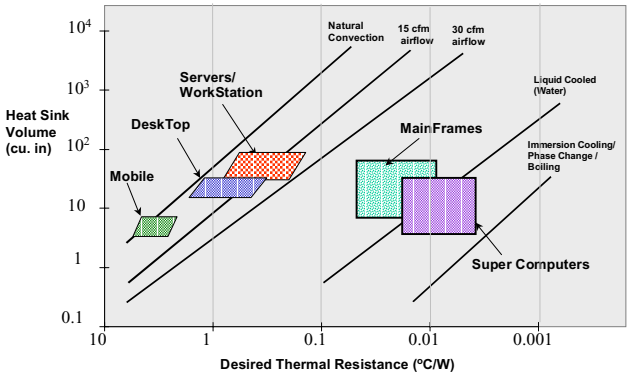


Figure 27: Thermal technology map

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

A Mechanism-Based Methodology for Processor Package Reliability Assessments

Nicholas P. Mencinger,
Assembly Technology Development Quality and Reliability, Intel Corp.

Index words: package, certification, reliability, methodology

ABSTRACT

Until recently, processor packages have been certified using a stress-based certification strategy. This approach, while easy to execute because the tests and end points are well defined, does not allow for an easy assessment of the fitness of the product under field applications. In fact, it doesn't require any knowledge of the end-user environment at all. The Assembly Technology Development Quality and Reliability (ATD Q&R) group has replaced this stress-based certification strategy with a mechanism-based strategy similar to one that was used extensively and successfully over many silicon technology generations. This paper describes the process by which the user environment was defined and discusses the application of this methodology to new processor package technologies.

INTRODUCTION

Intel processor packages up until now have been certified for reliability using a stress- or standards-based approach. This approach utilizes a fixed set of standard stresses of fixed durations to establish the field reliability of processor packages. It has been the standard means by which packages are certified throughout the industry. The standards are derived, or come directly from, the standards used to procure military electronics [1] and the methodology is described by Joint Electron Device Engineering Council JEDEC [2].

With the flip-chip (or C4 controlled collapse chip connect) generation of processor packages, questions concerning the use of military standards for certification were raised due to the extensive use of organic materials and the complexity of the design. It was assumed that the military standard stresses were either too severe or otherwise inappropriate for assessing these packages. These questions provided the impetus for reexamining the methodology.

There were three key motivators for reexamining Intel's processor package certification methodology. The first of these was the increasing segmentation of the processor market whereby packages are targeted for specific applications. This challenged the notion that one set of tests was adequate to address each market segment. For example, the duty cycles of server and notebook applications are significantly different and could result in overdesign or underdesign of the technology if both were required to meet the same test criteria. Moreover, because of market segmentation a processor is exposed to more environments, so an inflexible test suite may fail to adequately test specific field conditions.

The second motivator was the lifetime guarantee of military testing versus the customer's actual lifetime expectation. Given the rate of technological change, guaranteeing life to 100,000 hours seemed excessive. However, it wasn't clear what the customer lifetime expectation for a processor was since little data was available from the various market segments. In addition to an unclear lifetime expectation, it also was not clear what conditions did exist in the field that had to be met during that lifetime.

The final driver was the unclear link between the testing and the actual field use conditions, and moreover, what those conditions were. There are two conditions necessary for linking stress to life: a well defined user environment and physical models that links the environment to the accelerated stress test. In the existing stress-based methodology, neither the acceleration models nor the use conditions were clear; nor was it obvious that they were relevant to the now segmented processor market. Under these circumstances, highly accelerated tests appeared to be arbitrary, and in fact were unresponsive to new use environments.

These motivators were used to redefine ATD Q&R's certification methodology and align it to a mechanism- or knowledge-based approach. This approach is also described by JEDEC [3] but has been used at Intel only

reactively, i.e., for packaging when the stress-based requirements were not met. It generally has not been used in the industry due to its higher complexity and initial cost and the unclear definition of field lifetimes. The mechanism-based methodology requires that every failure mechanism be modeled against life with the appropriate physical model.

A related approach has been proposed by the high-density packaging user's group (HD-PUG) that attempts to define generic life models for various use conditions [4,5]. In this approach, the same acceleration models are applied but the model coefficients have been predetermined and are conservative. This approach, in essence, established more categories within a standards-based method, and because it was inherently conservative, did not achieve the full benefit of a mechanism-based approach.

This mechanism-based methodology was introduced on processor packages and is being considered by other packaging groups at Intel. The use condition methodology has already been adopted in silicon development as reliability limits have been reached in several areas. In general, the models have indicated that stress durations can be reduced, which has the direct benefit of saving package material and process costs and reducing time-to-information. It also allows greater flexibility in customizing reliability stress conditions so that they don't exceed the limits of the materials under test, which can generate test artifacts. Finally, since comprehensive models of the mechanisms are developed, they can be used to rapidly assess future extensions to the technology or to define when a technology will no longer meet the field requirements and needs to be replaced.

METHODOLOGY

In order to make the transition to a knowledge-based methodology, three things needed to be established: the lifetime expectations of the product, the environment in which the product was being used, and the tests necessary to simulate or accelerate that environment.

The lifetime expectations of the customers, both the OEM and the end user, were established by means of surveys. The surveys were conducted by market segment: server, performance PC, basic PC, and notebook; and also questioned whether the PC was for home or business use. From the survey data the expected lifetimes by market segment were established. Table 1 lists the survey areas broken into end-use conditions and OEM product qualification stress tests. The specific details of the survey are described elsewhere [6].

The surveys also asked specific questions concerning how the products are used in the field, hours of operation, on/off cycles, operating ambients, expected duty cycle, and storage conditions. In addition, data from processor junction temperature (Tj) and design temperatures were collected by segment for establishing the operating environment over the expected lifetime. These data were compiled by customers and used to develop use environments (or use conditions) by segment. The value or range chosen for each use condition captured approximately eighty percent of the total range of customer inputs for that condition and is therefore conservative, but not worst case. Extreme values, those beyond meteorologic possibility, were excluded.

Table 1: Survey questions

Use Conditions	Stress Tests
system lifetime	bake testing
cycles/Week (on/off) plus suspend/resumes	temperature cycling
on time hours per week	humidity testing
drop height	shock testing
use ambient range	drop testing
use humidity range	vibration testing
use/shipping vibration	
use/shipping shock	
storage temperature range	
storage humidity range	

Once the ambients were established from customer input, each was assigned to various accelerated test tools. Each accelerated test has an industry-accepted physical model that can link the test to the stress condition. These models were used to define the appropriate stress conditions and durations specific to the material set and failure mechanism being assessed.

Finally, in order to ensure customer acceptance of this methodology, it was necessary to develop support within the industry for the change through Sematech and to communicate the message to the customers. The Sematech Reliability Technical Advisory Board (RTAB) filled that role.

RESULTS

Lifetime Estimates

Figures 1 and 2 represent the summarized lifetime expectations for both OEMs and end users respectively. These data form the first key portion of a mechanism-

based certification methodology, namely, establishing the lifetime expectation.

Figure 1 is a summary of the expected lifetimes by market segment based upon input from the OEM survey. From the data it can be seen that, in general, there is little lifetime differentiation across the various market segments. The expected lifetime for all segments is in the seven year range, which was a surprising result. It was assumed that servers and notebooks would differ significantly in expected life, and while there was a small difference of less than one year noted, it was not deemed significant enough to warrant establishing a separate lifetime for notebooks.

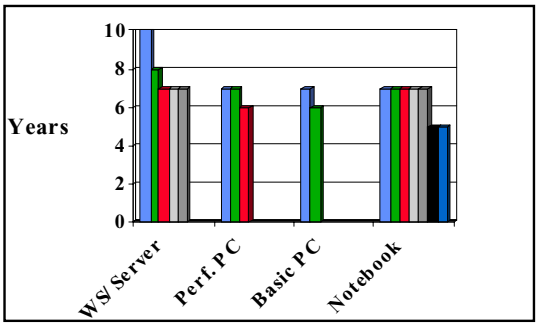


Figure 1: OEM Lifetime expectation by market segment; bars represent individual OEMs

Figure 2 is a lifetime model developed from the survey that asked end users the vintage of the processor in the computer they currently used. From that data it was possible to estimate the system replacement rate. In developing the model, an assumption was made that no replacement took place in the first three years of operation after which there was a twenty-eight percent replacement rate per year. The fitted model indicates that seventy-five percent of systems are replaced by year seven and ninety percent by year ten.

What is important to note here is that both the end user and OEMs expected lifetimes are very similar. Based on these data, the current expected lifetime for processor products was set at seven years with a small portion of the population remaining in service at ten years. Both numbers are important when modeling the wear out behavior of various failure mechanisms. Testing is taken out to a ten year equivalent to ensure that there are no catastrophic wear out mechanisms, that is, ones with narrow failure-rate distributions that result in a large proportion of the population failing over a short period of time.

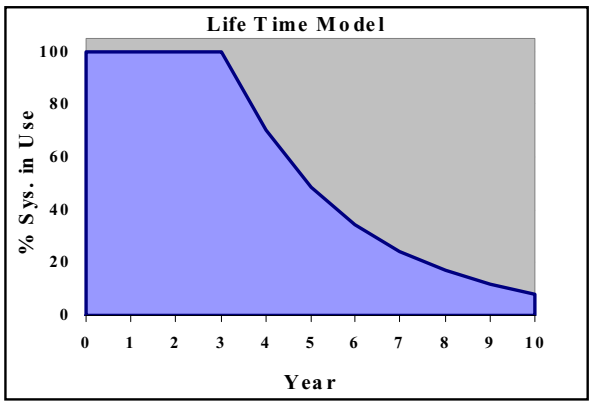


Figure 2: Lifetime model based on end-user input

Environment Definition

In Table 2, the input from the customer survey was analyzed and used to develop an environmental model. The environment takes into account the product exposure from the time it leaves Intel's manufacturing operation until the end of its useful life. Storage was broken out into two separate exposures covering two different environments. This was done since the accelerated test linkage for each category was different. The same was done for thermal and power cycles.

One of the surprising aspects of the input received was that there was little differentiation by market segment. In general, the stress on notebooks was found to be more severe than on stationary stystems for thermo-mechanical exposures such as thermal or power cycles. Notebooks were also more likely to have longer temperature-humidity-bias exposures due to short duty cycles and power save features. Servers, as expected, were at the other extreme with long duty cycles but fewer thermo-mechanical cycles. These differences were the exception however, and since processor products can be used in multiple market segments, the widest range is typically used when defining the stress conditions.

Table 2: Environmental exposures based on OEM input

Use Environment	Equivalent Use Condition by Market Segment*
moisture uptake in manufacturing	1 week out of bag
short duration extreme ambient temperature exposures during shipping and transport	-45 C to +75C for up to 24 hours
slow thermal cycles due to ambient changes or local heat sources (power supplies)	NB: 3000 cycles DT/Server/WS: 1500 cycles
fast processor On/off (to max. Tj) power cycles (including power save features)	Server/WS 3500 cycles DT/NB: 7500 cycles
operating air temperature range	10-35 C
ambient moisture during low-power state at operating voltages	62K hrs at 30C/85%RH
high operating temperature (Tj max at max. ambient)	62K Hrs
shipping vibration	Random
shipping shock or drop	0.5 m equivalent
operating vibration	Random
operating shock or drop	0.5 m equivalent
maximum sustained storage temperature	45 C Up to 1 year
minimum sustained storage temperature	-10 C Up to 1 year
socketings	15x max.
surface mount temperature exposures	3x @220C

* NB = notebook, DT = desktop, WS = workstation

Each of the environments in Table 2 were linked to the appropriate accelerated tests. The equivalent use condition from Table 2 and the lifetime model are used with acceleration models to define the accelerated test stress durations for each failure mechanism discovered during development. (The acceleration factor is the ratio of the failure rate in stress to the failure rate in the use environment.)

To ensure that future changes to the environment are comprehended and revisions made as appropriate, the survey will be repeated every two years with the OEMs. In addition, an internal review will be held for every new package technology being developed to address the detailed environment specific to that technology.

STRESS MODELS

The stress models used to link the environment and the accelerated tests were chosen because of their wide acceptance within the semiconductor packaging industry. The models in Table 3 are the baseline models being used for processor packages today and have been published in white papers by both Sematech's RTAB and Intel [7,8]. The methodology described here, however, does not preclude using other models when necessary. The Sematech models are only being used as a guide and will be modified according to the mechanism.

With the stress models' lifetime and environments defined, it becomes possible to use acceleration models to establish the expected field lifetime for every failure mechanism uncovered during development. The larger ramifications are 1) the failure models can be used to predict the impact of changes in the environment, for example, increased junction temperature effects on flip-chip (C4) joint resistance and 2) technology limits can be better defined and used for technology roadmap planning, for example, flip-chip bump pitch decreases limited by flip-chip metallurgy. With a stress-based methodology, these are not readily accomplished.

Failure Mechanism Modeling

Figure 3 gives the accelerations for several different flip-chip package thermal cycle failure mechanisms. In this example, the failure mechanisms are related through a power law relationship (Coffin-Manson) to an end-user environment of 1500 cycles with a ΔT of 40°C.

Table 3: Widely accepted acceleration models incorporated into the methodology

Mechanism	Model*
Temperature, Humidity Mechanisms	Peck's $TF = A_0 \times (a+bV) \times RH^{-N} \times \exp[E_a/kT]$ AF (ratio of TF values, Stress/use) = $[(a+bV_{Stress})/(a+bV_{Use})] \times (RH_{Stress}/RH_{use})^{-N} \times \exp[([E_a/k)(1/T_{Stress}-1/T_{use})]$
Thermal Effects	Arrhenius $TF = A_0 \times \exp[E_a/kT]$ AF (ratio of TF values, Bake/use) = $\exp[(E_a/k)(1/T_{Bake}-1/T_{use})]$
Temperature & Voltage Mechanisms	Eyring $TF = A_0 \times V^{-N} \times \exp[E_a/kT]$ AF (ratio of TF values, Stress/use) = $(V_{Stress}/V_{use})^{-N} \times \exp[(E_a/k)(1/T_{Stress}-1/T_{use})]$
Thermo-mechanical Mechanisms	Coffin-Manson Cycles to fail = $N_f = C_0 \times (\Delta T)^{-n}$ AF (ratio of N_f values, accelerated/use)= $N_{stress}/N_{use} = (\Delta T_{stress}/\Delta T_{use})^{-n}$

*TF = time to fail, AF = acceleration factor

Using the least accelerated mechanism with an exponent of 1.25 and a stress temperature ΔT of 150°C, (-25 to 125°C), it can be seen that approximately 400 cycles are required to simulate a lifetime. If a larger ΔT were chosen, fewer cycles would be required (300 for T/C B or –55 to 125°C). Typically, the most highly accelerated stress condition within the capability of the material set being stressed would be used to minimize the time to execute the test. By not being limited to the standard stress conditions, the advantage is that any stress condition can be chosen and the cycle count adjusted as required minimizing the risk of artifacts induced by overstressing.

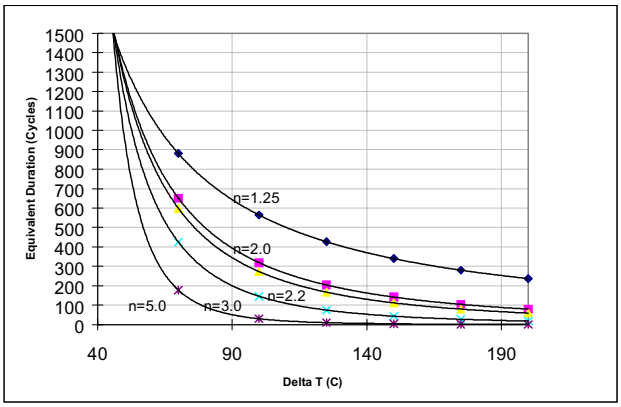


Figure 3: Thermal cycle counts at various ΔTs required to simulate 1500 environmental cycles with a ΔT of 40°C for various Coffin-Manson (power law) exponents

From this example, it can also be seen that highly accelerated mechanisms, those with exponents greater than three, only have to survive a few stress cycles. In this way, stress conditions and durations can be established for specific mechanisms while disregarding all other modes. In the stress-based-methodology, all mechanisms, regardless of acceleration, would be required to meet 1000 T/C B (ΔT = 180°C). From Figure 3, it can also be seen that requiring 1000 T/C B, as the stress-based model does, would result in designing for many lifetime multiples for highly accelerated mechanisms and would likely cost more than designed packages.

This model used a use-condition ΔT of 40°C. A key assumption was that for the flip-chip mechanisms, the only ΔT of consequence was that of the use environment. However, some thermo-mechanical failure mechanisms must be modeled from the neutral temperature of the package; that is, the temperature at which the stress in the package is zero and is typically near the molding or curing temperature. In this situation, the use ΔT will be considerably larger and will change the accelerations accordingly. Both situations should be modeled tofind the best fit, ensuring that the appropriate model is used. For a more thorough treatment of this point, see Reference 10.

In this second example, flip-chip interconnects were found to increase in resistance during bake. Flip-chip packages were baked at various temperatures and

periodic measurements of bump resistance were made. These measurements were used to establish the activation energy for this mechanism. The calculated activation energies for several different package types and vehicles ranged between 0.9 and 1.6eV with a best estimate of 1.3eV [11].

An analysis of DT system-operating temperatures was conducted and compared to actual measured system data and was found to be below 80°C [12]. The predicted mean and three sigma use-condition temperatures were used in an Arrhenius relationship with the failure-rate distribution and activation energy to estimate the time to 1% failure. From this analysis it was apparent that the lifetime at an upper operating temperature of 80°C was acceptable. Extrapolations of these results to higher operating temperatures indicated however that the lifetime needed to be increased. Thus, the acceleration models can be used for establishing a success criteria for the lifetime improvement team.

In the previous temperature cycle example multiple failure mechanisms were modeled. During temperature cycling, devices under test (DUT's) will fail for various failure mechanisms, each with a characteristic acceleration. The raw failure rates need to be transformed into use-condition fail rates using the appropriate acceleration factors. Mechanisms that fail early in stress may have long use condition lifetimes when transformed due to large acceleration factors. The transformed fail- rate models for each mechanism are summed up in a predicted cumulative lifetime fail-rate model, which is then compared against the lifetime expectation for the product and market segment.

ISSUES

There are a number of issues associated with the implementation of this methodology and to an extent these issues until now have precluded wide acceptance of this methodology. For the methodology to be successful, it requires that there be capability and capacity for running multiple stress conditions; three or more for each type of stress to be modeled. There is significant overhead for maintaining these facilities.

Several new thermal cycle conditions were defined to support the modeling of thermo-mechanical failure mechanisms adding to thermal cycle condition B and C already in use. The ranges of the new conditions overlap so that the failure mechanisms can be characterized both for Tneutral and Tstress. Table 4 lists the new conditions.

Table 4 : Thermal cycle ranges used for characterizing processor packages

Designator*	Temperature Range	ΔT
C	-65 to +150C	205C
B	-55 to +125C	180C
R	-25 to +125C	150C
Q	-25 to +100C	125C
T	0 to +125C	125C
S	0 to +100C	100C

*R, Q, T & S are internal to Intel and not published in JEDEC or Mil. Std. 883D documents.

Concomitant with multiple stress conditions is the need for the time and product volume necessary for developing acceleration models. The least accelerated legs of the testing can take months to accomplish and may require larger sample sizes due to low failure rates. The highly accelerated legs are then heavily relied upon for extrapolated estimates of the lifetime, which introduces significant uncertainty into the lifetime estimates. Since three or more conditions are run to develop a high-confidence model, three times the volume of product needs to be run to properly populate the stresses.

Highly accelerated tests run two major risks: that of introducing artifacts that wouldn't occur under less accelerated testing, and competing mechanisms. Artifacts require an increase in the failure analysis resources necessary to identify all the failure mechanisms. A comparison of failure modes between highly and less highly accelerated legs of a test sequence will identify which ones are artifacts, and these can be removed from consideration. More insidious is the risk of competing mechanisms that artificially depress a fail rate. Fab process 802 exhibited a corrosion mechanism during low- acceleration testing (85°C/85% RH) that was not seen in highly accelerated stress testing (HAST) [13]. The passivating effect of the highly accelerated test masked a failure mechanism that posed a significant field reliability risk. If the experimental design includes a sufficiently broad range of conditions, this risk should be minimized.

Customer and Industry Acceptance

Industry acceptance by semiconductor manufacturers was a key element in making this transition. Ensuring acceptance through the industry forums prevented competitors from using reliability methods as a competitive tool. One of the key forums was the

Sematech RTAB where the various manufacturers collaborated on the industry models and lifetime environmental exposures. The role of the RTAB was to reach consensus on the methodology, environment, and the method of communication to the customer base. The Sematech RTAB white paper was written and published jointly with other manufacturers and announced in a press release to industry trade journals [14,15]. Overall, acceptance by the industry has been excellent.

Intel's major customers were visited and given a presentation on the change to the methodology. The visits were timed to major product releases by market segment and served the purposes of informing the customer of the changes and of soliciting further feedback on the methods. None of the customers visited had any negative input, and most viewed the change positively. Customers whose input differed from the lifetime model presented typically asked where their specific input fell relative to other OEMs. Upon discussion of the data, none expressed significant concern that their input was different.

CONCLUSIONS

Based upon customer survey input, the expected lifetime for processor packages across all market segments was found to be seven years with a population still in service at ten years. Customer survey input was also used to define the user environment. More accurate processor package field lifetime estimates and risk assessments can be made based upon the lifetime and environmental models. This work has reestablished the link between the accelerated test methods and the field-use conditions. Through industry collaboration and customer communication, a new methodology for certification of processor packages based upon reliability failure mechanisms has been successfully implemented.

ACKNOWLEDGMENTS

I acknowledge the contributions of a number of people who provided invaluable help on this work, specifically Michael Dudash who developed and executed the survey; Steve Huber and Jack McCullen who worked with the industry groups on industry acceptance and who contributed to both Intel's internal and the Sematech RTAB white papers; Neal Mielke for his analysis of the resistance data used in the example; and Babak Sabi for his input on the lifetime model.

REFERENCES

1. Military Standard 883.D Notice 2, "Test Methods and Procedures for Microelectronics," 9/30/94.
2. JEDEC Standard 47, "Stress Test-Driven Reliability Qualifications of Integrated Circuits," July 1995.
3. JEDEC Standard 34, "Failure Mechanism Driven Reliability Qualification of Semiconductor Devices," March 1993.
4. HDP User Group, "Application Specific Semiconductor Device Qualification Methodology," *JEDEC Survey Ballot Submission*, 2/16/99.
5. The HDP User Group is an industry group formed to address application-specific package and product qualification methods. The executive board members include Amkor, Hewlett Packard, Lucent Nokia, Nortel Networks, and Sun.
6. Dudash, M., "Intel Customer Line Fallout DPM Goals and Use Condition Specification," *Intel Internal Specification 25GS0040*, 1999.
7. R. Blish, S. Huber, J. McCullen, N. Mencinger, "Use Condition Based Reliability Evaluation of New Package Technologies," *Sematech RTAB White Paper*, June 1999.
8. S. Huber, J. McCullen, N. Mencinger, *Intel White Paper*, 1999.
9. A . Lucero, N. Mencinger, R. Dias, "Predictive Reliability Modeling for C4 Lead Migration and Interconnect Bump Extrusion," *Intel Assembly Test Technology Journal*, 1999.
10. G . Shirley, J. McCullen, *IRPS Tutorial*.
11. Mielke, N. et al., unpublished bake resistance data, June 1999.
12. Mielke, N. et al., "WW21 P856 RelJET C4 Bake Resistance Update," *Intel Internal Document*, May 20, 1999.
13. Winters, K., "P802 THB Post Mortem," *Intel Internal Document*, March 1999.
14. The other members of the RTAB include IBM, TI, AMD, Conexant (Rockwell), Hewlett-Packard, Compaq, Lucent, and Motorola.
15. Lineback, J.R., "New Way to Qualify IC Packaging Could Speed Use of New Material," *Semiconductor Business News*, June 1999.

AUTHOR'S BIOGRAPHY

Nick Mencinger joined Intel's Assembly group in Santa Clara in 1980 after graduating from the University of Illinois with a B.S. degree in Ceramic Engineering. He transferred with that group to Chandler in 1984 eventually supporting various Assembly Technology

Development programs. In 1992, he joined the Corporate Quality Network supporting new processor package development programs. He currently comanages the ATD Q&R organization with Steve Huber. His e-mail is nicholas.p.mencinger@intel.com

Thermal Challenges During Microprocessor Testing

Pooya Tadayon, Sort Test Technology Development, Intel Corporation

Index words: thermal management, test, burn-in, electronics cooling, heat transfer

ABSTRACT

Thermal management of microprocessors during testing plays a key role in reducing cost while increasing yield and performance. Changes in packaging technology and the rapid increase in processor power and power density, however, are presenting unique thermal challenges that require innovative cooling solutions. The purpose of this paper is to inform the reader of the thermal challenges faced at Sort, Burn-In, and Class Test and to highlight some of the innovative solutions being developed to meet these challenges.

INTRODUCTION

There are three test steps in the manufacturing process (shown in Figure 1) where thermal management has an impact on the overall cost of a microprocessor. For example, adequate thermal control at Sort, where defective die are identified at wafer level, allows for the elimination of some downstream processes that ultimately result in considerable capital savings and faster time-to-market.

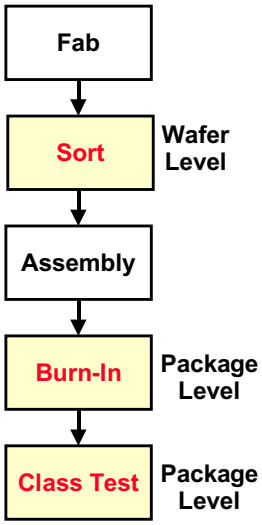


Figure 1: High-level manufacturing flow with key test steps highlighted in red

Similarly, it is important to control the die temperature, commonly referred to as junction temperature or T_j, during Burn-In (BI), where packaged units are stressed to accelerate early failures. Improving the thermal control at BI reduces the length of time that the devices need to be stressed, which results in less capital equipment expenditure and faster throughput time. An effective thermal solution at BI also leads to an increase in yield by allowing us to burn in devices that would otherwise go into thermal runaway. This is a phenomenon where the device draws more current as it gets hotter, which results in more self-heating and eventually leads to junction temperatures high enough to melt the package and possibly damage the equipment.

Finally, since the performance of an integrated circuit is highly dependent on the temperature of the device, it is of paramount importance to control the die temperature during Class Test as this is the step where we gauge the device performance at the component level. Any unnecessary increase in temperature during this test step will reduce the speed of the device by as much as 0.15% per degree celsius and decrease the yield of the fastest processors.

Based on the information provided above, it is clear that thermal management plays a very important role in the testing of microprocessors. Thus, it is necessary to control the die temperature during test where the goal is to gauge the device performance while keeping the test simple, efficient, and cost-effective. It is, however, extremely difficult to accurately control the temperature of the die since the power dissipation of logic devices can vary substantially during the test cycle (see Figure 2).

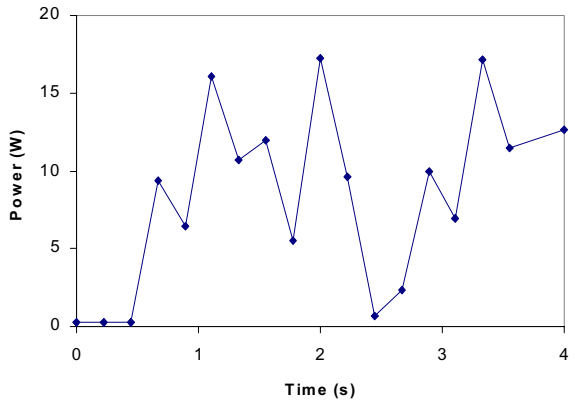


Figure 2: Typical power profile during the test cycle

This problem is exacerbated by non-uniform power distribution of highly integrated microprocessors, the introduction of flip-chip packages with an Integrated Heat Spreader (IHS), and the overall trend toward higher power and smaller features to maximize performance. Based on the extrapolation of historical trends shown in Figure 3, microprocessor power is expected to reach 200 W within the next five years with the average power density reaching values as high as 125 W/cm^2.

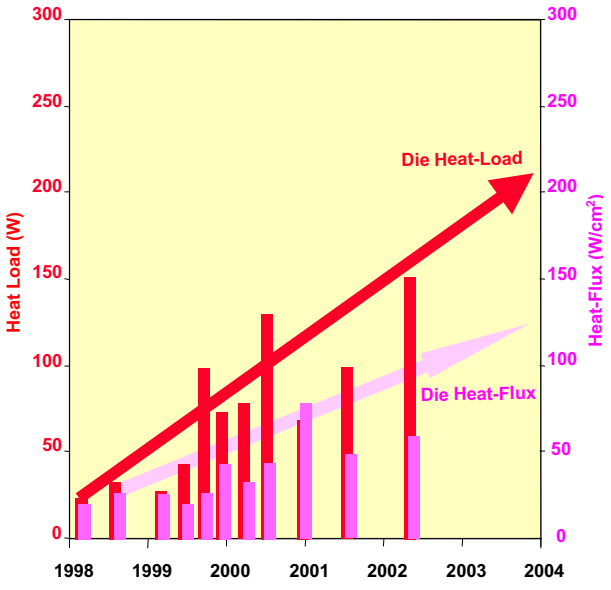


Figure 3: Microprocessor thermal roadmap based on extrapolation of historical trends

The industry trend towards flip-chip packages with an IHS is also presenting unique challenges at testing. The main purpose of the IHS is to reduce thermal gradients and enable the Original Equipment Manufacturer (OEM) thermal solution by providing a more uniform heat source and a more robust attach interface for the OEM heat sink. However, as shown in Figure 4, this increases the thermal resistance of the package and also eliminates direct access to the die, thus forcing us to control the junction temperature through the spreader.

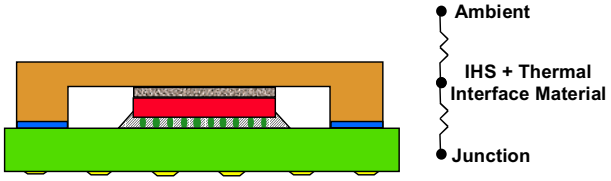


Figure 4: The thermal resistance stack-up for a flip-chip package with an IHS

Changes to the packaging and Si architecture, along with the need to supply the market with higher performance devices in a shorter time period, are challenging the existing thermal technologies at test and will require new and innovative solutions in order to help semiconductor manufacturers meet the market needs.

THERMAL CHALLENGES AT SORT

Wafer sort is the first step in the test process with its main purpose being to reduce assembly costs by identifying defective die at the wafer level so that these devices are not assembled.

Wafer sort is also the first step in the test process where thermal management becomes important. In the past, wafers were typically sorted at room temperature with little regard to thermal control of the Device Under Test (DUT). Today, however, wafers are sorted at cold temperatures, and the data are used to reduce test costs by eliminating several downstream test processes.

The idea behind cold testing is to identify and reject devices that fail at the low end of the specified operational temperature range. In previous generations of microprocessors, these failures were caught at Class Test where devices were tested at both hot and cold temperatures. In an effort to decrease the number of tests at Class Test and reduce costs, a method was developed to use Sort data to screen out devices that would otherwise fail at cold temperatures. This method, referred to as Cold Socket Elimination (CSE), currently requires that the DUT temperature be kept below 35 °C during Sort.

The current wafer probers use a thermal chuck to control the device temperature during Sort. The chuck is a Au-plated Al disc whose temperature is actively regulated to within ±1 C of the setpoint by an external chiller and heaters embedded underneath the disc.

The surface of the chuck contains several concentric rings with vacuum ports designed to hold down the wafers during testing. The contact between the wafer and the chuck, which plays a critical role in heat transfer, is enhanced during Sort as the probe card exerts up to 200 N of force on the die.

The thermal characteristics of the chuck have been evaluated using thermal test chips. The data, which are shown in Figure 5, indicate that with a setpoint of 0 °C, the chuck is capable of keeping the die temperature to about 25 °C for a steady state power of 70 W. This is well within the envelope of low-end products, which dissipate no more than 50 W during Sort. However, for future high-end products, which are expected to dissipate more than 100 W, the chuck will become a limiting factor as the die temperature will exceed the 35 °C T_j limit and put CSE at risk.

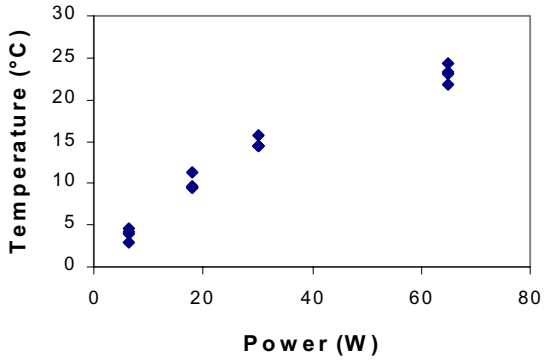


Figure 5: T_j as a function of power for wafers tested on a production prober under steady state power conditions

One quick solution to this problem is to lower the chuck setpoint temperature to below 0 °C. To illustrate this point, consider the definition of T_j

$$T_j = T_a + P \times \theta_{ja} \quad (1)$$

where T_a is the ambient or setpoint temperature, P is the device power, and θ_{ja} is the junction-to-ambient thermal resistance. Equation 1 indicates that for a given power and θ_{ja}, one can limit T_j by decreasing the setpoint temperature.

It has already been demonstrated that the existing probers can operate at -10 °C for an extended period of time without any problems. There is, however, a limit as to how much the setpoint temperature can be decreased. Lowering T_a below –10 °C will require expensive tool upgrades to enable the chiller to go down to such low temperatures and to prevent condensation inside the prober. In addition, reducing T_a may be practical for steady state conditions where there are little or no power fluctuations. As shown in Figure 2, however, there are considerable power fluctuations during the testing cycle, and lowering T_a could undercool the device during the low-power portions of the test and impact its reliability.

An alternate solution is to reduce θ_{ja} by improving the thermal contact between the wafer and the chuck through the use of a Thermal Interface Material (TIM). For example, there are currently probers on the market that use water as the TIM and can reportedly dissipate up to several hundred watts of power while maintaining an acceptable junction temperature. There are, of course, a myriad of problems associated with using a liquid interface such as tool complexity, maintenance, reliability, and safety. Liquid interfaces also tend to stain and/or leave a residue on the backside of the wafer that can create problems in the subsequent assembly and test steps. Alternatively, it is possible to reduce the wafer-to-chuck thermal resistance by using a dry TIM such as thermally conductive flexible foils that are readily available on the market. Some of these materials have been shown to reduce the thermal resistance, and hence T_j rise, by up to 30%.

Another option is to optimize the chuck material and its manufacturing process. Recent data show that replacing Al with Cu, which has a ~2X higher thermal conductivity, and polishing the chuck surface to reduce surface roughness improves the thermal performance of the chuck by more than 50%. The combination of lowering the setpoint temperature, changing the chuck material, polishing the chuck surface, and using a TIM may yield sufficient margin to meet future product requirements.

Thermal control is one of the main focus areas as Intel plans its transition from 200mm to 300mm wafers. Based on the roadmap shown in Figure 3, the 300mm probers may need to dissipate up to 200 W while keeping T_j below 35 °C. Future probers may use some form of direct air impingement on the die or active thermal control in order to achieve better thermal control.

THERMAL CHALLENGES AT BURN-IN

Burn-In is a batch process where up to a thousand assembled units are simultaneously stressed at elevated temperatures and voltages in order to accelerate latent reliability defects and processing problems to failure. The key challenge at BI is to keep the BI time low in order to decrease throughput time and minimize equipment and processing costs.

BI time is a function of many variables including the outgoing failure rate, yield, die size, voltage, and junction temperature. The outgoing failure rate, or DPM goal, is defined by corporate policy while yield and die size are

process and product attributes, respectively. The two variables that can be manipulated from a manufacturing process standpoint are voltage and T_j.

Since voltage yields a higher acceleration factor than temperature, it is desirable to burn in devices at the highest possible voltage in order to maximize the acceleration factor and minimize BI time. The maximum BI voltage has historically been defined as 1.4X use voltage and cannot be increased further without damaging the device.

BI time can also be minimized by ensuring that T_j is as high as possible but below the functionality limit for all the units within the BI oven; any variation in T_j translates into longer BI times. To illustrate this point, consider Figure 6 which shows the calculated T_j distribution in the current generation and Next-Generation Burn-In (NGBI) ovens. Since BI time is a function of the median T_j, devices in the NGBI chamber that have a tighter distribution and a higher median T_j will have a lower BI time. In this particular simulation, the median or BI T_j in the NGBI chamber is about 14 °C higher than in the current BI system. According to the plot in Figure 7, this 14 °C increase in BI temperature results in about a three hour decrease in BI time.

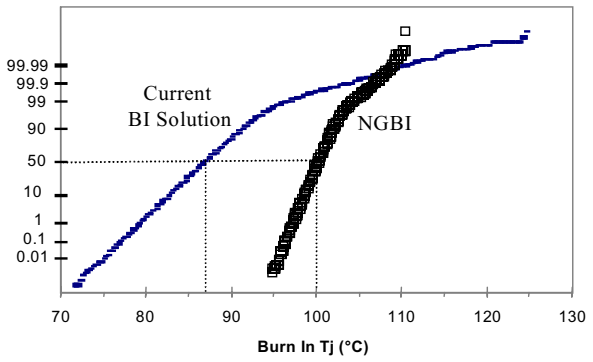


Figure 6: Calculated *Tj* distribution in the current and next-generation BI ovens

In addition to reducing the BI time, tightening the T_j distribution also helps increase yield by enabling burn in of units that are at the tail end of the distribution. Due to concerns over thermal runaway and device functionality, the BI T_j cannot exceed the maximum functionality limit. If we assume that the maximum BI T_j in the simulation shown in Figure 6 is 110 °C, then the units at the tail end of the distribution that have a T_j greater than 110 °C

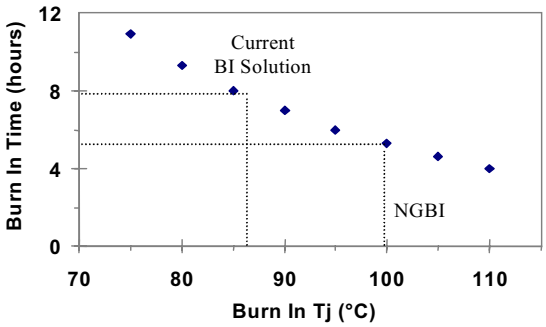


Figure 7: Calculated BI time as a function of BI T_j

would have to be scrapped. This translates to a ~0.1% yield loss with the current BI solution. The improved thermal capability of the NGBI system, however, allows these devices to be burned in, thus resulting in an increase in yield.

It is clear that the only way to maximize BI temperature without shifting part of the distribution over the max T_j limit is to reduce the T_j variation. To better understand the sources of variation in T_j, we refer the reader to Equation 1 where T_j is expressed in terms of T_a , P , and θ_{ja}. Each of these variables has an inherent variation associated with it that contributes to the overall T_j variation.

The variation in T_a is a function of BI hardware technology and can be minimized at the expense of module complexity and cost. For high-power devices, however, the second term in Equation 1 is the dominant source of T_j variation, and further hardware improvements to reduce T_a variation do not significantly affect the T_j distribution.

Power variations are mainly a function of the wafer manufacturing process. Since BI power is a function of transistor and gate leakage, any variation in the silicon fabrication process that affects transistor and gate leakage will directly translate into a variation in BI power. Unfortunately, there is not much that can be done from a test process development point of view to reduce these power variations. It is, however, possible to minimize the effects of power variations by reducing θ_{ja}.

Besides the absolute value of θ_{ja}, the variation in the thermal resistance is also a key factor. Large variations will amplify the power variations and lead to a broader T_j distribution. Thus, minimizing θ_{ja} and its variation in the BI environment is a major challenge as up to a thousand units are being processed simultaneously in a single oven.

In addition to maintaining a tight T_j distribution, another key challenge in the BI environment is the ability to dump

the total heat dissipated by the units into the environment. This has generally not been a problem for previous generation processors whose BI power was under 10 W, thus requiring the BI oven to dissipate less than 10 kW of heat. As transistor features shrink and leakage increases, however, the BI power is expected to exceed 250 W per DUT. This means that the BI oven must be capable of dissipating more than 250 kW in order to enable burn in of several hundred to a thousand devices. The alternative to not meeting this capacity requirement is to purchase extra ovens, which will take up additional factory floor space and increase the overall cost of the process.

Figure 8 shows a schematic diagram of the air-cooled BI oven currently being used in manufacturing. The thermal solution consists of a BI socket with an integrated anodized Al heat sink that makes contact with the die when a device is placed inside the socket. Forced-air convection is then used to remove the heat from the heat sinks and an air-to-air heat exchanger is used to dump the heat into the environment.

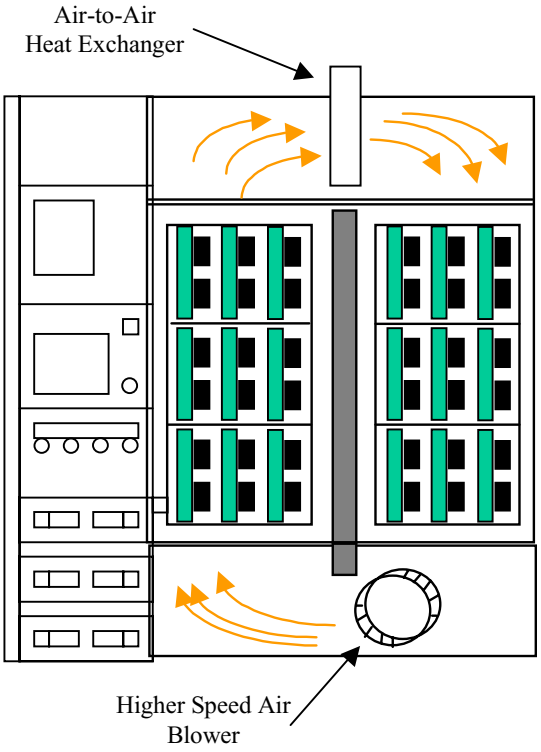


Figure 8: Schematic diagram of a typical air cooled BI oven with the BI boards and BI sockets displayed in green and black, respectively

This module is capable of dissipating 6-8 kW for typical setpoint temperatures of 65-80 °C and can achieve a θ_{ja} of 4.6 °C/W with a standard deviation of 0.7 °C/W for a typical 1 cm^2 device without an IHS. This is sufficient to meet the requirements of previous-generation microprocessors. Future-generation products, however, will require a θ_{ja} of less than 1 °C/W and a much higher dissipation capability in order to meet the expected BI time targets.

One approach taken to extend the capabilities of the existing system was to increase the height of the heat sink in order to increase the surface area of the fins. Due to space constraints, however, the oven had to be depopulated by every other slot so that the heat sinks would not come in contact with adjacent burn-in boards. This configuration yielded a θ_{ja} of 2.4 °C/W with a standard deviation of 0.3 °C/W but resulted in a 50% decrease in oven capacity which, for most High-Volume Manufacturing (HVM) products, is an unacceptable tradeoff.

Other schemes to improve the module capability include retrofitting the ovens with a larger blower and an air-to-liquid heat exchanger. The larger blower increased the air flow within the chamber and improved θ_{ja} by up to 30%, while the addition of an air-to-liquid heat exchanger improved the overall power dissipation capability by more than 2X. These module enhancements, however, are point solutions that provide near term capability and it is obvious that a new system is needed to meet long-term product requirements.

The limitations imposed by the current BI solution prompted the development of the NGBI system. The key features of NGBI are that it reduces the ambient temperature variations by a factor of two, increases the system-level power dissipation capability by as much as a factor of three, and uses a novel solution to decrease θ_{ja} by nearly an order of magnitude.

The ambient temperature control and the system-level power dissipation of the NGBI chamber is significantly better because it uses a liquid medium instead of air. The system employs a Cu heat sink, or a button, that is cooled by forced-liquid convection. The fluidics system is designed to ensure uniform flow across each button, thus reducing ambient temperature variations due to uneven flow. In addition, the high-heat capacity of liquids and the use of a liquid-to-liquid heat exchanger allows the system to dissipate more than 50 kW per chamber.

What makes NGBI special is the use of a eutectic alloy interface to improve the thermal contact between the die and the button. The alloy liquefies at elevated temperatures and makes nearly perfect contact with the die and the button. The advantage of the alloy interface is that it is a liquid metal that has very high thermal conductivity and yields a θ_{ja} of ~0.5 °C/W with a standard deviation of less than 0.1 °C/W. The disadvantages of

this solution are that it tends to leave a residue on the device and that it is still a laboratory solution that has not been proven to function in an HVM environment. The key challenge for the development team is to optimize the recipe and the process to enable the use of this interface material in the factories.

Changes to the packaging architecture, however, will continue to challenge even the best thermal solutions. As shown in Figure 4, the addition of an IHS to flip-chip packages increases the total thermal resistance, which directly impacts the BI process. The plot in Figure 9 show that the addition of an IHS increases θ_{ja} and its variability by nearly 2X, which ultimately leads to longer BI times and possibly lower yields.

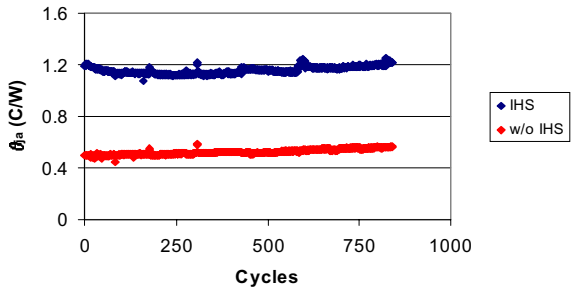


Figure 9: Thermal impedance of alloy for devices with and without an IHS

The extendibility of the NGBI module for future generations has been a topic of interest in light of the rapidly increasing BI power due to aggressive junction scaling. Estimates show that BI power could very well exceed 250 W in the next five years. Thermal management of a thousand devices dissipating 250 W each is a daunting, yet unique, challenge that requires extensive ingenuity and engineering.

Unless major changes are made within the Si to limit transistor and gate oxide leakage, future products will continue to challenge the existing BI solution even further. Future BI systems may employ more direct forms of liquid cooling such as liquid immersion, which has been previously used in the industry to burn in high-power devices. There is, of course, a number of issues with such a solution including the safety of the highly expensive dielectric fluid used as the coolant and the general concern over having a hot liquid bath in a factory environment.

A more promising solution is single DUT active thermal control where it is possible to achieve very tight T_j distributions by individually regulating the temperature of each DUT. Although much more attractive than immersion cooling from a safety standpoint, such a solution introduces a high level of hardware and software complexity that presents a unique set of challenges and risks.

It is widely agreed that we are pushing the limits of the current BI technologies and that innovative solutions such as liquid immersion, jet impingement, or active cooling may be needed to meet future product requirements. One of the key challenges in this endeavor is to develop a solution that not only meets the technical requirements but is also cost effective and suitable for an HVM factory.

THERMAL CHALLENGES AT CLASS TEST

One of the final steps in the manufacturing process is Class Test where the device undergoes a final series of tests to validate functionality and determine the speed of the part. One of the key requirements at Class Test is to ensure that the device is tested at or above the use temperature specified to the customer and at the same time keep T_j below the maximum reliability temperature. Thus, temperature control at Class Test is of paramount importance since it is critical to minimize T_j rise above the use, or setpoint, temperature in order to increase the yield of top-speed bins.

To illustrate this point, consider the simulation in Figure 10, which shows the T_j rise profile for the same device tested under two different conditions. The simulation shows that the T_j rise during the speed-binning portion of the test can be reduced by ~20 °C by simply using a heat sink with direct air impingement. This reduction in T_j rise translates to a ~3% increase in processor speed, which ultimately leads to an increase in the yield of high-speed devices.

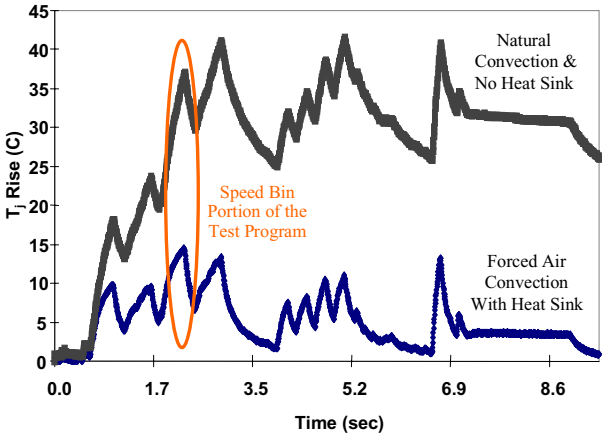


Figure 10: Simulation showing the impact of improved thermal control on T_j rise during Class Test

Intel's high-power products have continuously challenged the thermal control technology used during Class Test. The thermal solutions used in previous generations did not employ any heat sinking solutions and relied on natural convection to keep the devices cool. This method worked well for Plastic Land Grid Array (PLGA) packages that had a large thermal mass due to the Cu heat slug that was bonded to the die (see Figure 11).

With the introduction of Organic Land Grid Array (OLGA) packages, which have a very low thermal mass, thermal management became more of a concern as these devices had a ~5X higher T_j rise during Class Test than their predecessors. This problem was solved by integrating a Ni-plated Al heat sink into the test chuck in order to replicate the heat sinking capabilities of the PLGA packages. This solution improved the overall thermal capabilities of the handler and reduced T_j rise by nearly a factor of ten. In addition, direct-air impingement to the heat sink was used to further improve the thermal capabilities of the system so that it could handle even higher power devices.

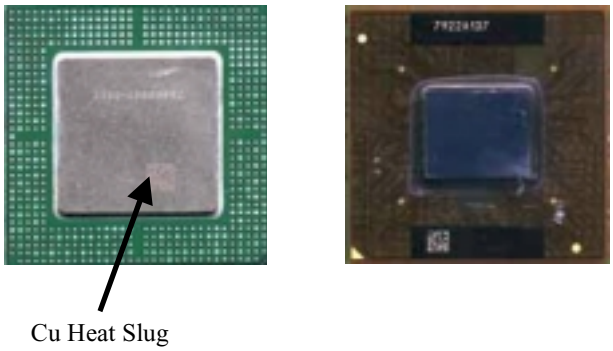


Figure 11: Physical differences between PLGA (left) and OLGA (right) packages

The latest migration to new microprocessor architectures and highly integrated devices has led to an increase in total power over previous-generation processors. As a result, a new thermal solution was needed in order to ensure that Class Test was not the limiting factor in the race for higher speed processors.

A major advance in the current-generation thermal solution is the use of a liquid interface between the device and the heat sink to reduce the thermal resistance and minimize T_j rise during test. In addition, the Au-plated Cu heat sink is cooled by liquid impingement, which is far more efficient and effective than air impingement. Data show that devices tested on handlers equipped with this technology are on average 10 MHz faster than if they were tested on the previous-generation equipment. Although the liquid interface presented a lot of technical and manufacturing challenges, it was necessary in order to meet the expected performance needs.

The continuous increase in power and the addition of an IHS to flip-chip packages, however, is once again challenging the thermal solution at Class Test. As discussed in detail previously, the key issues with the IHS are that it adds another thermal resistance to the stackup and it requires that we control T_j without direct access to the die. The addition of an IHS increases the total thermal resistance by up to 2X, which translates directly to a higher T_j rise during test.

In addition, as processors become more integrated, the impact of non-uniform heating during Class Test also becomes significant. For example, the local or peak power density for a given device could be as much as an order of magnitude higher than the average power density. This non-uniform power distribution leads to temperature gradients and makes it nearly impossible to maintain a constant T_j across the die. Simulations show that even with today's thermal control technology, the temperature in the local hot spot regions will easily exceed the maximum reliability temperature and increase the risk of damaging the device.

One short-term solution to address some of the thermal issues at Class Test is to lower the setpoint and use non-speed or non-temperature sensitive patterns to warm-up the die temperature to that of the use condition before speed-block patterns are tested. T_j rise could be reduced by minimizing the power difference between the speed-block patterns and "warm-up" patterns.

The long-term solution is to develop a new thermal solution for Class Test. The core technology of today's thermal solution is the water-based liquid interface, which is limited by its critical heat flux (CHF) and cannot handle devices with a power density greater than ~100 W/cm^2. Additionally, the liquid-cooled heat sink is

approaching the limits of passive thermal control. An active thermal solution, with the ability to cool hot spots at various locations on the die, is needed to meet the challenges set forth by the next generation of microprocessors.

Figure 12 shows recent data comparing the existing passive solution against a prototype system where active thermal control was employed to cool a 50 W processor with an IHS. The temperature profiles clearly indicate the superior performance of the active control solution, even in the case where an Interface Fluid (IF) was not used. The key challenge with this technology is developing a robust feedback mechanism that is compatible with a wide range of test equipment and products.

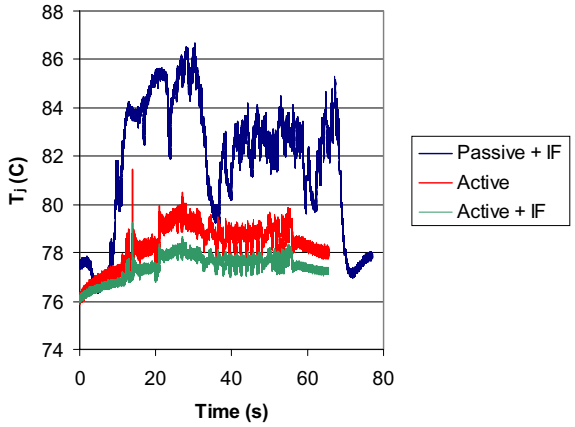


Figure 12: Data showing the impact of active thermal control on T_j

CONCLUSION

The intent of this paper has been to describe to the reader the importance of thermal management during microprocessor testing and the key thermal challenges at Sort, BI, and Class Test along with some of the solutions that are being developed to meet future product requirements.

The most difficult challenges are at BI where the temperature of up to a thousand units must be controlled simultaneously in order to minimize BI time. This requirement, along with the rapid increase in BI power, is driving for solutions that are capable of providing near zero θ_{ja} with the ability to dissipate large quantities of heat.

Thermal control at Class Test is important since the performance of a processor is a function of temperature, and lack of an adequate thermal solution directly impacts the company's competitive edge and revenues. New and innovative solutions are needed to deal with the rapid increase in power, changes in packaging technology, and the market need for faster products.

Finally, the less stringent requirements at Sort ease the thermal challenges and do not require that we develop exotic high-risk technologies. In fact, it is important to recognize that there is a limit to how good the thermal control needs to be at each test step so that excessive resources are not spent on developing high-risk technologies that are not HVM compatible.

ACKNOWLEDGMENTS

I thank Hongfei Yan, Jason Glumbik, Philip R. Martin, and Arun Krishnamoorthy for providing the data necessary to complete this paper. I also thank Mike Mayberry, Frank Monzon, and Ravi Mahajan for their generous comments and suggestions.

AUTHOR'S BIOGRAPHY

Pooya Tadayon is an Integration Engineer with Intel's Sort Test Technology Development group in Portland, Oregon. His current focus is investigating new thermal technologies that can be integrated into the manufacturing environment. Dr. Tadayon received a B.S. degree in Chemistry, Biochemistry, and Biology from the University of Washington and a Ph.D. in Physical Chemistry from Oregon State University. His e-mail address is pooya.tadayon@intel.com.