

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.

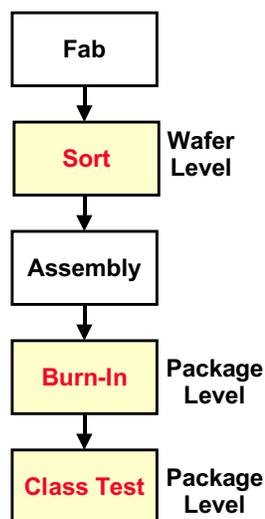


Figure 1: High-level manufacturing flow with key test steps highlighted in red

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).

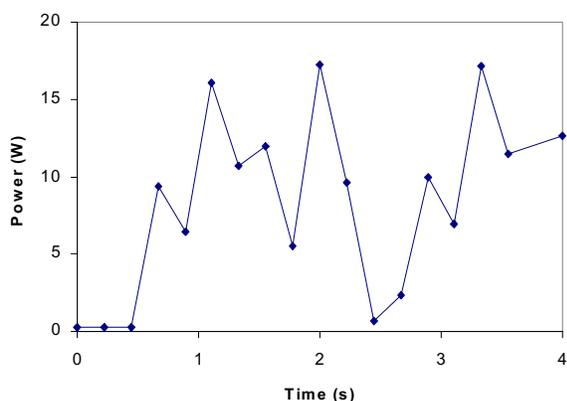


Figure 2: Typical power profile during the test cycle

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/cm².

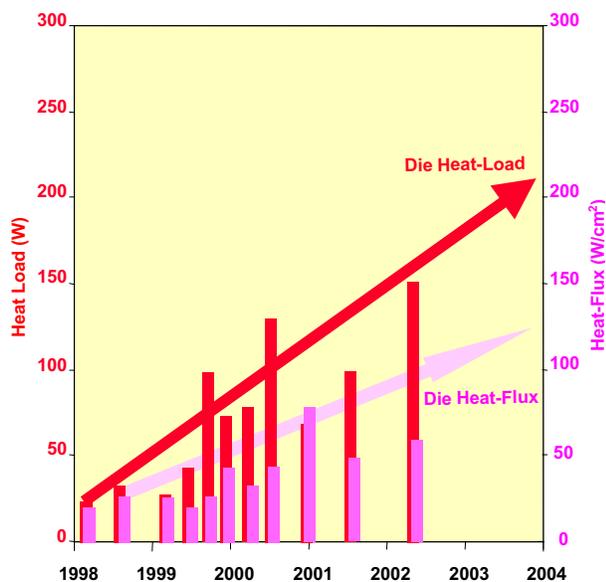


Figure 3: Microprocessor thermal roadmap based on extrapolation of historical trends

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.

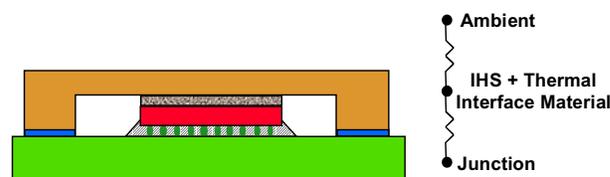


Figure 4: The thermal resistance stack-up for a flip-chip package with an IHS

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.

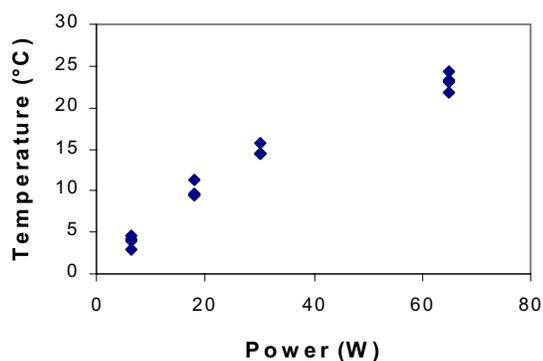


Figure 5: T_j as a function of power for wafers tested on a production prober under steady state power conditions

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.

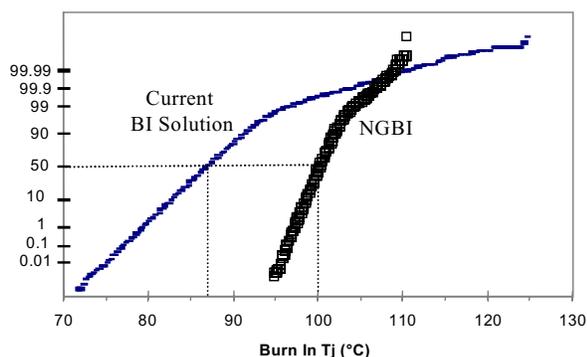


Figure 6: Calculated T_j distribution in the current and next-generation BI ovens

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

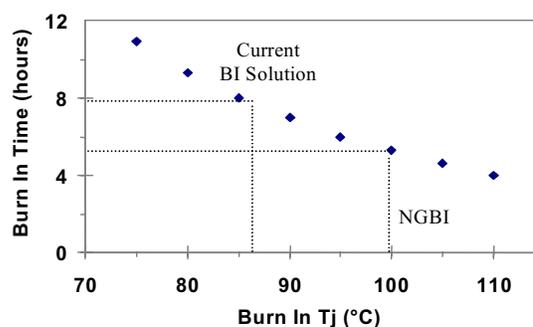


Figure 7: Calculated BI time as a function of BI T_j

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.

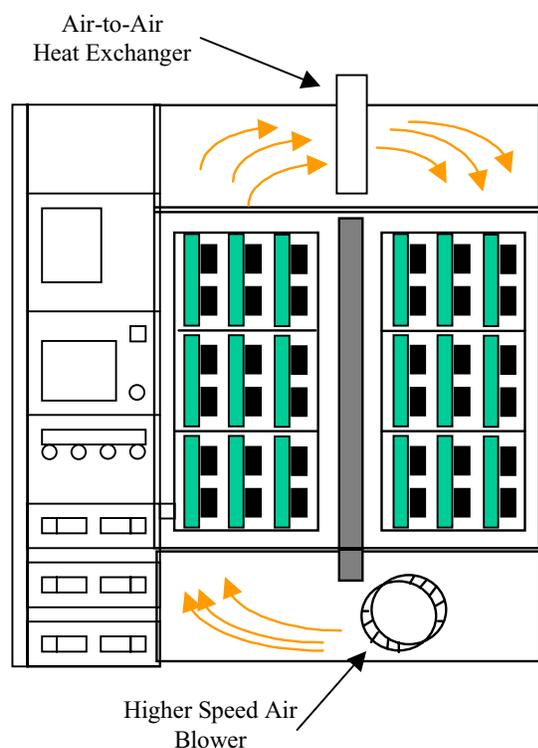


Figure 8: Schematic diagram of a typical air cooled BI oven with the BI boards and BI sockets displayed in green and black, respectively

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 cm² 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 shows 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.

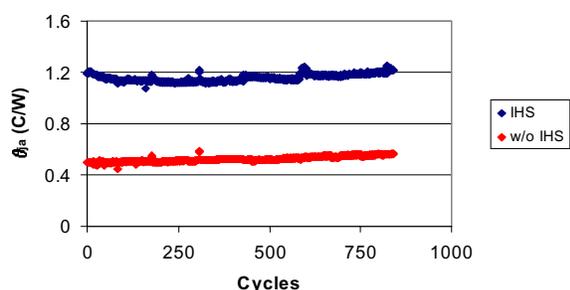


Figure 9: Thermal impedance of alloy for devices with and without an IHS

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 $\sim 20^\circ\text{C}$ by simply using a heat sink with direct air impingement. This reduction in T_j rise translates to a $\sim 3\%$ increase in processor speed, which ultimately leads to an increase in the yield of high-speed devices.

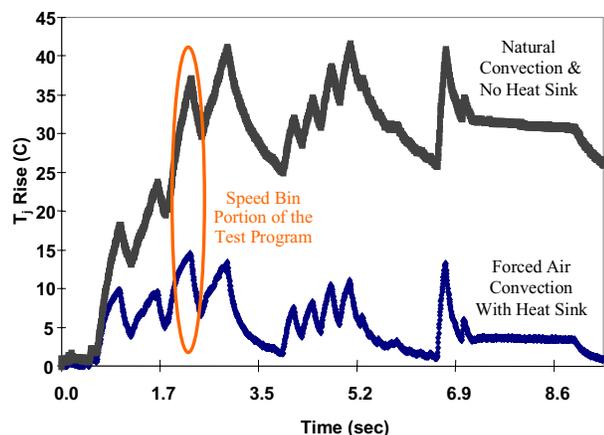


Figure 10: Simulation showing the impact of improved thermal control on T_j rise during Class Test

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.

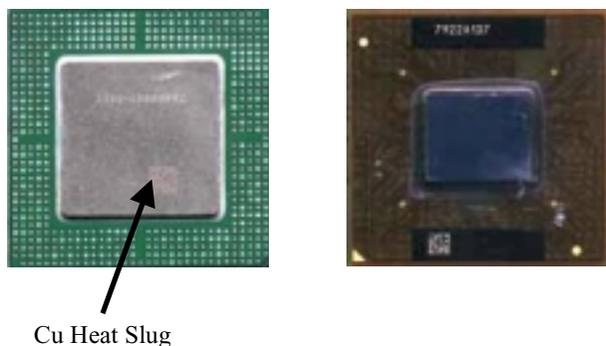


Figure 11: Physical differences between PLGA (left) and OLGA (right) packages

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/cm². 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.

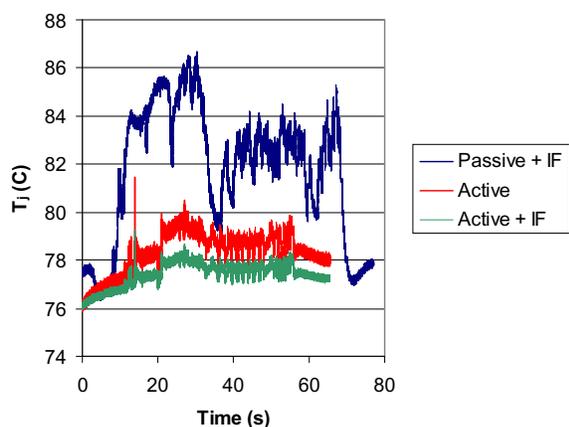


Figure 12: Data showing the impact of active thermal control on T_j

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.