High Productivity Thermo-Compression Flip Chip Bonding

Tom Colosimo, Horst Clauberg, Evan Galipeau, Matthew B. Wasserman, Michael Schmidt-Lange and Bob Chylak

Kulicke and Soffa Ind., Inc.
1005 Virginia Dr.
Fort Washington, PA 19034, USA
Ph: +1-215-784-6733
Email: hclauberg@kns.com

Abstract

Advancements in electronic packaging performance and cost have historically been driven by higher integration primarily provided by fab shrinks that has followed the well-known Moore's law. However, due to the tremendous and continuously increasing cost of building new fabs, the performance/cost improvements achieved via node shrinks are negated. This leaves packaging innovation as the vehicle to achieve future cost-performance improvements. This has initiated a More-than-Moore idea that has led to vigorous R&D in packaging. Advanced packages which employ ultra-fine pitch flip chip technology for chip-to-substrate, chip-to-chip, or chip-to-interposer for the first level interconnect have been developed as an answer to obtaining higher performance. However, the costs are too high as compared to traditional wire bonding. The status today is that the fundamental technical hurdles of manufacturing the new advanced packages have been solved, but cost reduction and yield improvements have to be addressed for large-scale adoption into high volume manufacturing.

In traditional flip chip assembly silicon chips are tacked onto a substrate and then the solder joints are melted and mass reflowed in an oven. This mass reflow technique is troublesome as the pitch of the solder bumps become finer. This is due to the large differences in the thermal expansion coefficient of the die and the substrate, which creates stress at the solder joints and warpage of the package when the die and substrate are heated and cooled together. To mitigate and resolve this issue, thermo-compression bonders have been developed which locally reflow the solder without subjecting the entire substrate to the heating and cooling cycle. This requires that the bondhead undergo heating past the melting point of solder and then cooling down to a low enough temperature to pick the next die from the wafer that is mounted to tape. Machines in the market today can accomplish this temperature cycle in 7 to 15 seconds. This is substantially slower than the standard flip chip process which leads to high cost and is delaying the introduction of these new packages.

This paper shows a flip chip bonder with a new heating and cooling concept that will radically improve the productivity of thermo-compression bonding. Data and productivity cycles from this new bond head with heating rates of over 200ºC/sec and cooling of faster than 100ºC/sec are revealed. Experimental results are shown of exceptional temperature accuracy across the die of 5ºC throughout the cycle and better than 3 ºC at the final heating stage. The high speed thermo-compression bonds are analyzed and the efficacy of the new concept is proven. Excellent temperature uniformity while heating rapidly is an absolute necessity for enabling good solder joints in a fast process. Without good temperature uniformity, additional dwell times need to be incorporated to allow heat to flow to all of the joints, negating any benefits from rapid heating.
Whereas the current state-of-that-art is often to program temperature in steps, this bonder can be commanded and accurately follows more complex temperature profiles with great accuracy. Examples of how this profiling can be used to enhance the uniformity and integrity of the joints with non-conductive pastes, film, and without underfill along with the associated productivity improvements will be shown. Tests that show portability across platforms that will lead to set up time and yield improvements and are identified and quantified. Additionally new ideas for materials and equipment development to further enhance productivity and yield are explored.

Key words
3D packaging, Advanced packaging, Cu-pillar, Flip-chip, Thermocompression

-

I. Introduction

Thermocompression (TC) flip-chip bonding is seen as one of the key enablers for next-generation electronic packages. Standard flip-chip bonding cannot be used for I/O pitch below about 100µm, mainly because it cannot control the flow of solder sufficiently well and because the large difference in the coefficient of thermal expansion (CTE) between silicon and organic substrates result in large stresses and warpage after the mass reflow process. These problems are exacerbated for thin die and die stacks. Thermocompression bonding resolves these issues by minimizing the temperature to which the organic substrate it exposed and much more closely controlling the bondline thickness.

Adoption of TC bonding for high volume production has nonetheless been somewhat slow because of limitation of the first-generation bonders that have been available until now. The throughput has been limited to only a few hundred units per hour (UPH), bringing the cost-of-ownership of these complicated, high-cost machines into a range that is unacceptable for most applications. K&S has now developed a TC bonder that greatly improves the cost-of-ownership by enabling throughput for real processes in the 1000-2000 UPH range. These high throughput values are a result of thoroughly understanding the process requirements, incorporating this understanding into the bonder design and exceptionally high heating and cooling rates of the bondhead while maintaining temperature uniformity and micron-level control over X, Y, Z and tilt.

The K&S bonder used for this study has a unique split-axis architecture that intrinsically enables better process control. The bondhead moves only in Y and Z, while the substrate stage moves on both a coarse X-axis for indexing and a sub-micron X-axis for micron-level accuracy. The split axis means that there is only one hand-over for the die (from the picker to the bondhead), while still allowing the bonder to apply the thermocompression force directly in-line with the center of the die. Most other bonder architectures do not apply the force in-line with the die, which may cause die skidding under high forces, and most also have multiple hand-overs of the die.

II. Thermocompression Process Overview

Thermocompression processes can be divided into a few distinct process types, each with their own advantages and disadvantages. These are: Non-Conductive Paste (NCP), Underfill Film (UF) applied to the die, and dip fluxing of the die followed by TC bonding without any pre-applied underfill. Each process type has distinct benefits and disadvantages.

Bond quality, process stability and throughput factor into which process may be best for any given application. Since the pre-applied underfills (paste and film) are substantially cured during the TC process, both distribute any stresses between the die and the substrate over the entire die area immediately after the bonding process is complete. In contrast, TC processes without pre-applied underfill causes all of the stress to be concentrated onto the Cu pillars until the die is underfilled, usually by capillary underfill in a subsequent process step. These stresses, of course, are still much lower than for a standard flip-chip process, but more careful die and substrate designs are still needed to manage these stresses. Capillary underfill of large die with a narrow bondline can also be quite challenging.

Non-conductive paste and underfill films, however, present their own challenges. Especially with paste, there is a substantial probability of too much paste wetting out beyond the die edges and then contaminating the top of the die and the bonding tool for die that are thinner than about 100µm. Even with perfect control of dispensed volume, variations in bondline thickness can easily reduce the volume of material under the die and cause excessive wet-out. Since Cu pillar heights can vary across a wafer, bondline thickness will also vary and excessive wet-out can be a problem even with perfect control of dispense volume. Even underfill films will encounter some problems with excessive wet-out for very thin die because of the Cu pillar height variation.

Process throughput and the associated cost-of -ownership is, of course, also an absolutely critical factor in determining whether a new bonding technology like thermocompression bonding can be commercially successful. A certain amount of throughput loss over standard flip-chip bonding must be accepted when going to

a thermocompression process where a solder joint is formed right on the die bonder. The benefit being that finer pitches at lower stress and lower warpage are enabled and a few down-stream processes like flux cleaning or capillary underfill might be eliminated.

The main process time penalty for TC bonding comes from the thermal excursions that the bondhead needs to cycle through for every placement. Even at 200ºC/sec up and 100ºC/sec down, a cycle of 180ºC → 260ºC → 180ºC will take a minimum of 1.2 sec. Adding a very fast cycle time of 1.2 sec and considering that some of cooling can be performed during the bonder motions, the limit on UPH for a bonder with two TC bonding heads will be just above 3000. Additional processes like dip fluxing, the need to cool to lower temperatures (as for underfill films) or the need to add dwell times to cure pre-applied underfills will subtract from this upper limit UPH, in some cases substantially. Selection and optimization of a TC process for highest throughput should minimize thermal excursions. Excellent control over temperature uniformity and place tool position in X, Y and Z are also crucial for enabling good processes at high temperature ramp rates. With a well-designed bonder and careful attention to the bonding process throughputs far above the current industry status of 300 – 400 UPH (dual head) are achievable. Current low UPH values are not intrinsic limitations of the TC bonding process, but limitations imposed by the equipment

We will now analyze each component of throughput in more detail. The TC bonding process consists of the following phases, not all of which are applicable to all process flows:
Die Transfer – handing the die from the pick tool to the place tool
Dip fluxing – if applicable, dip Cu pillar tips into flux
Motions to alignment position
Alignment – up and down-looking camera aligns the die to the substrate
Motion to seek height
Seek – slower motion to approach and impact the substrate
Bonding – multiple process steps to distribute the paste/film, cure it, melt the solder and then cool to freeze the solder – again, not all are needed for all processes
Lift off and return to the die transfer position

With an eye toward minimizing the process times and temperature excursions, we can now examine all of these stages. Since the upper temperature is more or less fixed to around 280ºC by the need to melt the solder, it is mainly process steps that require the bonding head to be at low temperature that force large temperature excursions.

Die transfer – Processes involving underfill film require the lowest bondhead temperature at die transfer. During transfer, the underfill film is in direct contact with the pick tool. If the bondhead is too hot when it touches the backside of the die, the film can melt and deform ir even adhere to the pick tool. Underfill film limits the transfer temperature to about 80ºC. For the other two processes types, the maximum temperature is limited to about 200ºC by the pick tool.

Flux dipping - This only applies to the flux dip process where die are dipped into flux and then placed without any underfill. The temperature is limited to about 100ºC to avoid evaporating or activating the flux too early. The flux dipping process itself adds approximately 400ms to the process time.

Alignment – In the case of underfill film, one cannot let the film become too liquid. In the case of dip flux, one does not want to start evaporating the flux or activate it too soon. In both cases the temperature is limited to the range of 100 –150ºC, depending on materials. For NCP, the temperature can be around 170ºC, again material dependent.

Seek and impact – For pre-applied underfill (paste or film), the seek motion and impact needs to be carefully controlled to allow flow and filling of voids. For NCP the starting seek height is relatively high because the motion needs to be well controlled starting at the height of the dispensed bead. The temperature during seek needs to be high enough for easy flow of the paste, but not so high as to start curing it. For the dip flux process, the temperature just needs to remain low enough for the flux to not evaporate.

Bonding process – For the dip flux process, the temperature ramp simply consists of heating to above the melting point and cooling back down to freeze the solder. The bonding process in this case can be very fast, but requires excellent control over Z position, since the die will be free floating once the solder melts. The boding process for an underfill process will often start at a lower temperature and possibly have some additional time built into it to allow for sufficient flow and curing of the adhesive.

III. Heater Performance

A. Heater Temperature profiles

The heater and its heating and cooling performance is, of course, the heart of the TC bonder. Figure 1 shows the surface temperature of the K&S bonder place tool during a temperature cycle from 160ºC to 280ºC and back to 160ºC with a 1 sec dwell time at 280ºC. The temperature of the tool was measured using a temperature-sensing infrared camera. The actual heating rate of the surface of the tool,

not just the commanded heating rate or the temperature at the heater elements, is 200ºC/sec. The cooling rate at the tool surface averages to about 135ºC/sec for cooling from 280ºC to 160ºC. Since the bonder uses room temperature air for cooling, the instantaneous rate is of course a little higher at the top of the heat cycle than near the lower end. Any specifications of heating or cooling rate should always identify the range over which the performance is achieved.

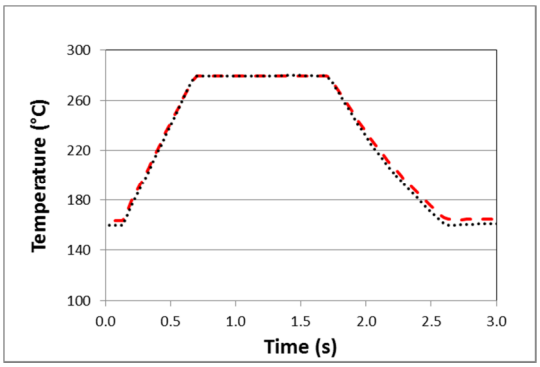

Figure 1: 160-280-160ºC Temperature cycle. Tool surface (red dashes), internal temperature sensor (black dots)

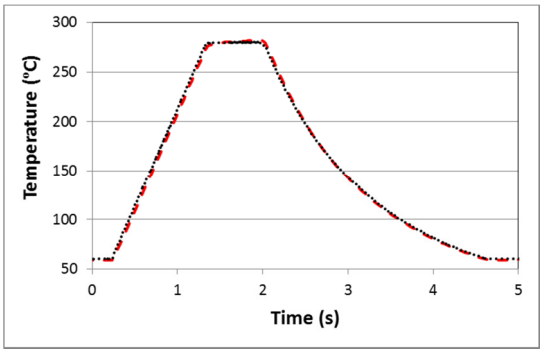

Figure 2: 60-280-60ºC Temperature cycle. Tool surface (red dashes), internal temperature sensor (black dots)

Figure 2 shows a similar temperature measurements for a 60ºC to 280ºC and back temperature cycle that might be needed for a underfill film process. In both cases one should note the close tracking between the internal temperature sensor that is used to control the heater in a feedback loop and the tool surface itself. This level of tight tracking enables the excellent process control needed for fast processes. The two temperature cycles, even with a 1 sec dwell time, require only about 2.5 and 4 seconds.

In Figure 3 we now examine the effect of placing a die on the heater surface. In this case, the heat flow will be somewhat retarded by the interface between the die and the tool and by the thickness of the die itself. One can see that during both heating and cooling, the temperature of the die surface lags behind the internal temperature sensor by only about 160 to 170 ms with the 125um thick die used for this experiment. The temperature profile in this experiment used a 500ms dwell time to allow the bottom die surface to come within about 5ºC of the command upper temperature of 280ºC.

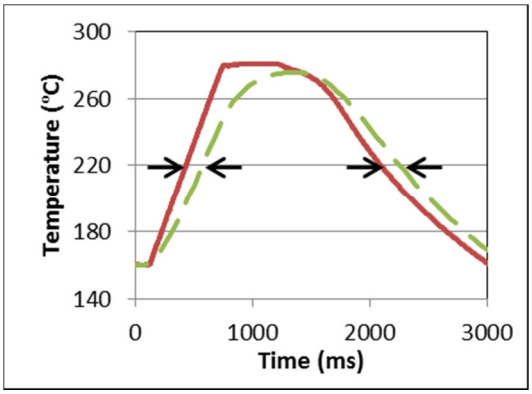

Figure 3: Temperature comparison of the internal temperature sensor (red solid) and the surface of a 125µm thick die on the place tool (green dashes)

Depending on the type of process being run, the tool could lift off the die well before the 160ºC is achieved during cooling. For a dip flux process, the tool could lift off around 200ºC, or about 2 sec into the heating/cooling cycle. The rest of the cooling could be performed with the place tool in the air and a substantial part of the cooling can occur during the return path of the bond head to get the next die. In an underfill film or non-conducting paste process, the tool could lift off near the top of the curve.

B. Temperature uniformity

High heating and cooling rates are only one part of the equation for achieving good bond quality in a fast TC bonding process. Temperature uniformity is just as critical. Nothing is gained if the center of the tool heats up fast, but the corners lag behind or vice versa. Through careful modeling and innovative design and manufacturing techniques, K&S has been able to develop a heater with exceptional temperature uniformity during 200ºC/sec heating and >100ºC/sec cooling.

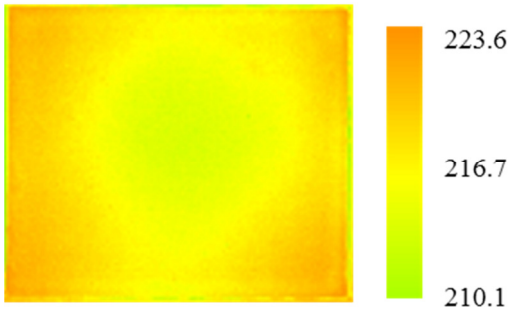

Figure 4: Temperature uniformity of a 125µm thick die near 220ºC in the temperature profile of Fig. 6.

Figures 4 and 5 show thermal images of an approximately 9 x 10mm, 125µm thick die on a place tool at the midpoint and the top of the temperature excursion in Figure 6. At the top of the temperature cycle, the temperatures are all within a ±2ºC range. Even at the midpoint, during the 200ºC/sec ramp, the temperature is within a ±5ºC. This exceptionally good temperature uniformity is also demonstrated in Figure 6, which shows overlapped temperature profiles of the four corners and the center during the temperature cycle.

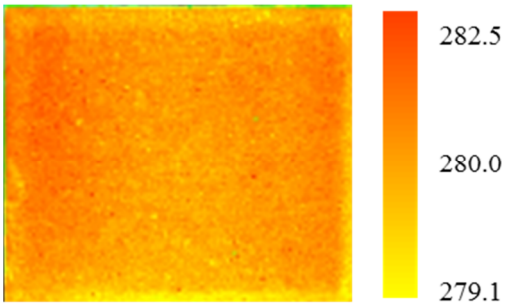

Figure 5: Temperature uniformity of a 125µm thick die near 280ºC

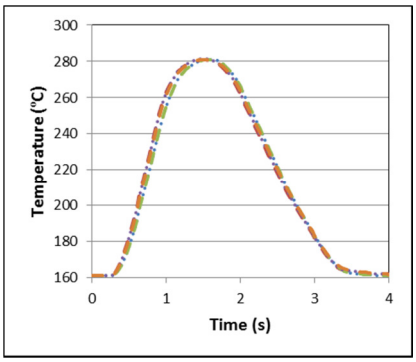

Figure 6: Temperature of 1.5 x 1.5 mm areas in the corners and at the center of an approximately 9 x 10 mm die.

IV. Importance of High Heating and Cooling Rates

In this section we focus on the reasons why high heating and cooling rates are also necessary for high quality solder joints. TC bonding relies on high thermal gradients in the Z direction and it is precisely these high Z temperature gradients that allow packages with lower internal stress than a standard FC process. The challenge is to achieve high thermal gradients in Z while heating all solder joints to the same temperature for nearly the same amount of time.

A. Die stacking without re-melting layers

One application in which very high Z thermal gradients are highly important is the stacking of thin dies without any underfill. As mentioned previously, the use of pre-applied underfill is challenging because of the problem of excessive wet-out and the associated potential for place tool contamination. Of course, when stacking die without underfill, it is imperative that as a die is added to the stack, the solder in the previously bonded lower layer does not melt again. If it does, control of the bondline thickness is lost. Figure 7 shows a stack of four thin dies with good solder joints and bondline thickness for all layers. The process could only be achieved with temperature ramps of near 200ºC/sec. All bonding attempts using about 100ºC/sec heating caused some re-melting of lower layers whenever good solder joints were formed in the top layer.

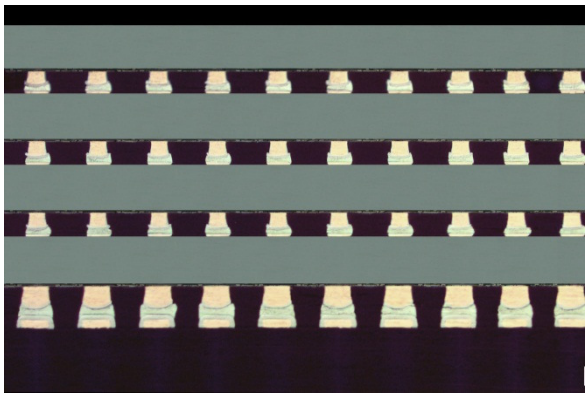

Figure 7: Four-die stack of thin die without underfill

B. Long lead effect

Substrate leads of varying length present different thermal sinks to the Cu pillar that is being bonded to them. Long leads have a larger thermal mass and more contact to the cool substrate. A simplified finite-element model (Figure 8) with Cu traces of 100µm and 500µm length was developed to better understand this process. The 75µm thick die is heated from the top, while the bottom of the 200µm thick substrate is held at 70ºC. The Cu pillars have a diameter of 25ºm and are 50ºm apart. To demonstrate how high heating rates provide better control over solder flow, the top of the die was heated at either 100ºC/sec or 200ºC/sec from a starting temperature of 70ºC. The cooling rate was left at 100ºC/sec in both cases.

The case of 100ºC/sec heating to 270ºC, 100ºC/sec cooling and a dwell time of 50ms at 270ºC is shown in Figure 9(a) and the same case except heating at 200ºC/sec in Figure 9(b). The figures show the temperature at the top of the die and in the solder joints to the short and the long leads. The first striking feature of the model is that even for perfectly uniform heating from the top, two leads only 50µm apart can have a temperature difference of 20ºC simply because of a difference in their length. Also very interesting is that for any given temperature at the top of the die, the temperature difference between the short lead and the long lead is the same, whether the top of the die is heated at 100ºC/sec or 200ºC/sec. This indicates that the dynamics of

heat flow in the die and leads are much faster than the timescale of the heat input at the top of the die.

But closer comparison of the two cases reveals a critical difference (Table I) in the length of time that the solder joints are above the melting point, taken to be 217ºC. When heating at 100ºC/sec, the long lead is above melting for 360ms, whereas the short lead is above melting for 780ms. This means that the solder on the short lead has an addition 420ms in which it might flow and potentially wick out of the solder joint. In the case of heating at 200ºC/sec, the time difference in cut to 300ms. Of course, slower cooling rates would have exactly the same effect of increasing the time difference for which each joint is molten.

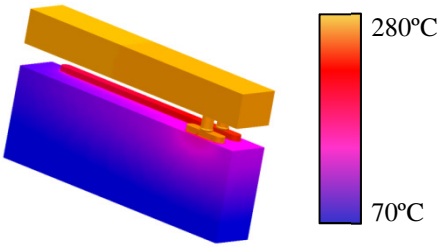

Figure 8: Temperature distribution in long and short traces during a TC bonding cycle.

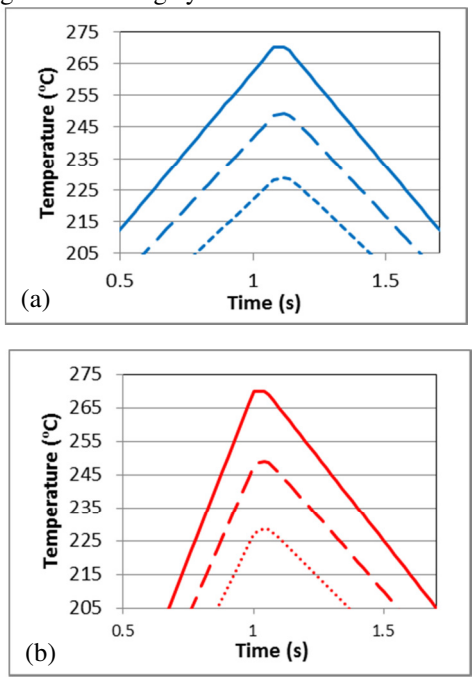

Figure 9: Temperature profiles for the top of the die(solid), the short trace(dashes) and the long trace (dots) for the model in Figure 8 under 100ºC/sec (a) and 200ºC/sec (b) heating.

The model clearly illustrates how critically important fast heating and cooling are for controlling solder flow. We have observed exactly this effect when bonding our test die. The second lead from the right in Figure 10 is much longer than the others. With insufficient heating, the solder does not wet the long lead, while all the short leads are well wetted. Only with carefully tuning of the temperature profile to provide enough heat to this trace without allowing too much time for solder to flow for the short leads could the good results of Figure 10(top) be achieved. The effect of too much heating is seen in Figure 11. Here the solder has flowed too much and rather than taking on a rounded profile, it has assumed an undesirable hour-glass shape.

Table I: Time above the solder melting point for the short and long traces in the model of Figure 8.

	100C/sec heating	200C/sec heating
Short trace	780ms	580ms
Long trace	360ms	280ms
Difference	420ms	300ms

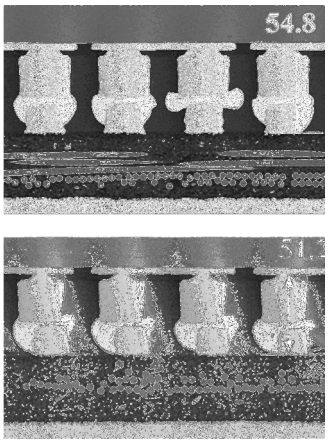

Figure 10: Cold joint to the long lead with insufficient heating (top) and good joint to the same lead with an optimized temperature profile

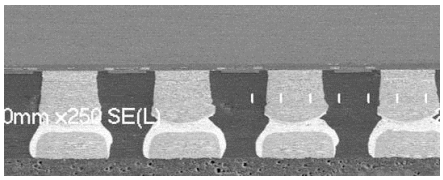

Figure 11: Solder can wick out of the joint if too much heat is applied

V. Conclusion

High heating and cooling rates while maintaining very uniform temperature are critically important to achieving

not only high throughput in a thermocompression bonding process, but also for creating high-quality thermocompression joints. The close interaction between die, substrate, underfill material (or lack thereof) and bonding process has been demonstrated. Successful thermocompression bonding can only be achieved through a combination of the right bonder, the right materials and thorough understanding of the process.