Low Tg Underfill: The Good, The Bad, and The Ugly

Craig Hillman, Nathan Blattau, Gil Sharon

Underfill, in some respects, has been an impressive technological advance for component packaging and follows similar historical trends for many innovations. In some respects, the real revolution was the concept of 'flipping' over the die face down and using solderable connections to create a high density array of connections between the integrated circuit and the substrate. This feat is attributed to IBM through the introduction of its controlled collapse chip connection (C4) technology in the 360 Mainframe System in 1964. The C4 was not quite the flip chip we know today, as it had solid copper balls covered in solder. Subsequent work by IBM, DEC, and Motorola eventually drove the adoption of an industry standard flip chip design with high lead solder bumps (97Pb/3Sn).

Amusing thought: Most references acknowledge IBM as the creator of the technology, but the terminology 'flip chip' comes from DEC.

The development of flip chip packaging in 1961 is amazing when one considers that the invention of the integrated circuit only occurred two years before (Jack Kilby of Texas Instruments and Robert Noyce of Fairchild filed their patents in 1959). In fact, it was so early, the initial flip chip packaging was applied to transistor devices, not integrated circuits. And yet while integrated circuit technology went from 10 transistors to 1 million transistors over the next 30 years, flip chip technology, by relative comparison, stagnated. Flip chip packages, while offering substantial density and electrical performance improvements, remained a niche product limited to die below a certain size (distance to neutral point) while wire bonds, the technology it was designed to replaced, soared in popularity.

Why the discrepancy? Need, cost, and knowledge. The need was based on the reality that silicon flip devices have a very low coefficient of thermal expansion (approximately 2.5 ppm/C). To survive the assembly process and any temperature variation in the field, flip chip devices had to be assembled to ceramic substrates. The cost was that ceramic substrates can be 2X to 10X the cost of organic-based substrates (typically FR-4 or Bis-aleimide Triazine (BT)). And there was a lack of knowledge on how to mitigate this restriction. The result was that flip chip technology remained at levels far below 1% of the annual integrated circuit production.

Which takes us back to the original focus of this article: underfill. Like many technologies, the adoption of flip chip was limited until the introduction of an enabler technology. Examples such as LEDs, touch-screen, and composites have all gone through this process. In some respects, the electronic industry was lucky as the enabling technology does not emerge until 40 to 50 years after the initial invention. In the case of underfill, just as with flip chip, there were several companies investigating the technology in the late 1980's. However, Hitachi is often recognized as the first company to publish and patent the concept that is most aligned with the underfill we know today.

With the introduction of underfill, the market for flip chip technology increases rapidly as flip chips can now be bonded to low-cost organic substrates. Through an extensive amount of testing and material design of experiments performed at IBM, the preferred properties of the underfill is a high modulus, low CTE material. Through several iterations, this ends up being primarily an epoxy with an elevated glass transition temperature (between 125C to 150C) and highly filled with silica (glass) beads with a 'sufficient' amount of shrinkage (I'll explain this a little later).

Through tweaking of the polymer chemistry and the size distribution of the filler particles, underfill manufacturers are able to greatly improve manufacturability, expand capillary function, and reduce cure time and temperature.

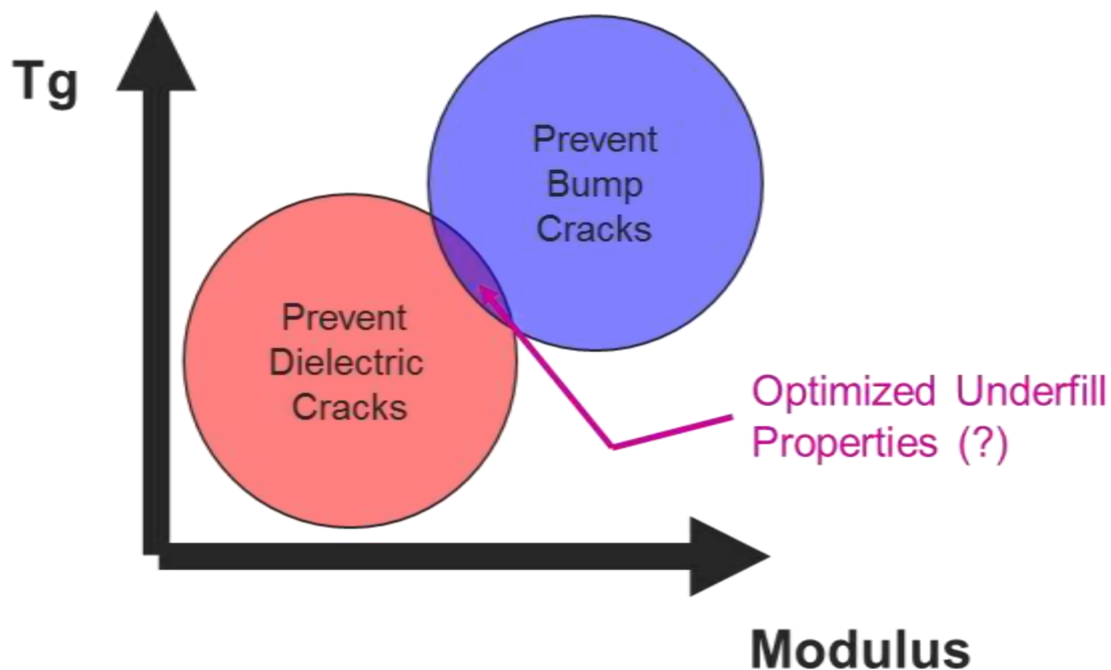
So far, so good, right? And it potentially gets even better. Several authors' stated that since the cost of wirebonding is I/O *dependent* while the cost of solder bumping is I/O *independent*, there would be a cross-over point where it would be cheaper to go from wire bond to flip chip. As it usual with these kinds of projections, the reality turned out to be more complicated, as the predicted threshold went from 100 I/O, to 300 I/O, to 500 I/O, etc. (you get the picture). A more up to date, fascinating, and nuanced analysis was recently completed by Chet Palesko and Jan Vardaman. As of 2010, the cost cross-over seemed to be around 1000 I/O for gold wire bonds and not until 2000 I/O for copper wire bonds. [Reminds me of the constant prediction that solid state drives will be cheaper than disc drives. These predictions often fail to consider the human element, which is people working on the existing technology do not like to give up their jobs willingly.]

But, there are other reasons to migrate to flip chip packaging (size, electrical performance, thermal performance) and the watershed moment for the technology is likely when Intel releases the Pentium III in 1999, which is the first flip chip CPU on an organic substrate. This is in the middle of a rapid surge in flip chip packaging, where flip chip as a percentage of integrated circuit packaging goes from less than 1% to almost 10% over two decades.

And then, as usual, the market threw a curve ball. Actually, it threw two curve balls (apologies for the sport reference for those readers that do not have baseball in their home country). The first market adjustment was the increasing demand for mobile electronics. As described by Ken Gileo, initial attempts to leverage existing underfill materials for flip chip on board (FCOB) in mobile devices were an abject failure because of a dramatic change in environments. Here was an application that could care less about temperature or power cycling, unlike server applications (IBM / Intel), but was extremely sensitive to mechanical loads (specifically bending/handling and mechanical shock).

The second curve ball was the use of porous SiO_2 as a low-K dielectric starting with the 90nm technology node devices. While porous SiO_2 as an elegant solution to the challenging problem of signal delay (due to an ever increasing RC time constant), it simply kicked the can down the road to the packaging level. The result of the weaker dielectric (what else would you expect when you add voids to a material?) was the introduction of the 'white bump' failure mechanism, where the low-K dielectric would crack after underfill cure (depending on bump geometry, under bump metallurgy, die size, etc.).

The solution, to both problems, was to 'soften' (reduce the elastic modulus) the underfill. The most straight forward way to change this mechanical property and not radically change other elements of the material was to lower the glass transition temperature (Tg). However, the Tg could not be reduced too dramatically or its original value would be eliminated. This conundrum lead to the 'optimum' low Tg underfill which prevent mechanical failures, eliminate white bump failures, and still maintain temperature cycling robustness. It was almost too good to be true. And it was…



Now, to start, we should probably define exactly what low Tg underfill is (the what). From my perspective, low Tg underfill is defined as a glass-filled epoxy with a glass transition temperature below the maximum use temperature of the device. Note (actually, a very important note): Not below the maximum storage temperature or the maximum test temperature or the maximum junction temperature. Only when it is below the maximum use temperature.

The introduction of low-Tg underfill into the marketplace was followed by rapid acceptance by most of the major manufacturers of flip chip ball grid arrays (FCBGAs). Central processing units (CPUs), graphic processing units (GPUs), and field programmable gate arrays (FPGAs) fabricated with low-K dielectric (the transition was between 45nm and 90nm process node depending on the technology) quickly integrated the new material set into their packaging construction. White bump issues faded to the background and the new material passed all JESD22-A104 temperature cycling tests with flying colors. The low Tg underfill was so good that it became literally impossible to get flip chip bumps to fail under temperature cycling. It was literally bullet-proof! And that should have been the first indication that something was very wrong.

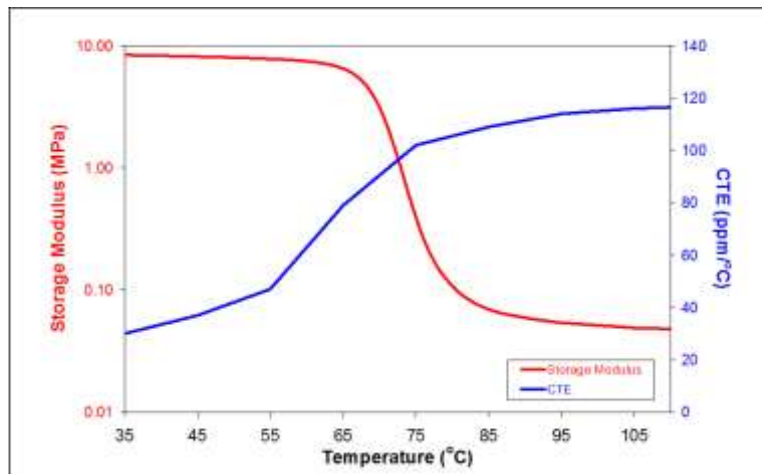
The problems first became apparent in between 2005 and 2007. Numerous vendors started reporting extremely high failure rates on some products, with some indications of either a 60% failure rate or repair rates exceeding 100% (the average unit came back for repair more than once). Initially, the failures were difficult to diagnosis and some organizations struggled with identifying root-cause for several months. The reason for this dilemma was a combination of the nature of the failure mode, the architecture of flip chip packaging, and the random location of the failure site.

What has happening, as we now know, was the solder bumps under the flip chip device were failing due to thermomechanical fatigue. However, due to the array nature of flip chips (this is a problem with ball grid array packaging as well), the nearest neighbors will force the crack closed. This results in a failure mode that is intermittent (sometimes there, sometimes not). This intermittency was so well known, that some users of a particular gaming system that was experiencing flip chip cracking would wrap the unit in towels and force it to overheat. This would increase the intermittency duration long enough for them to resell the gaming system back to retailers.

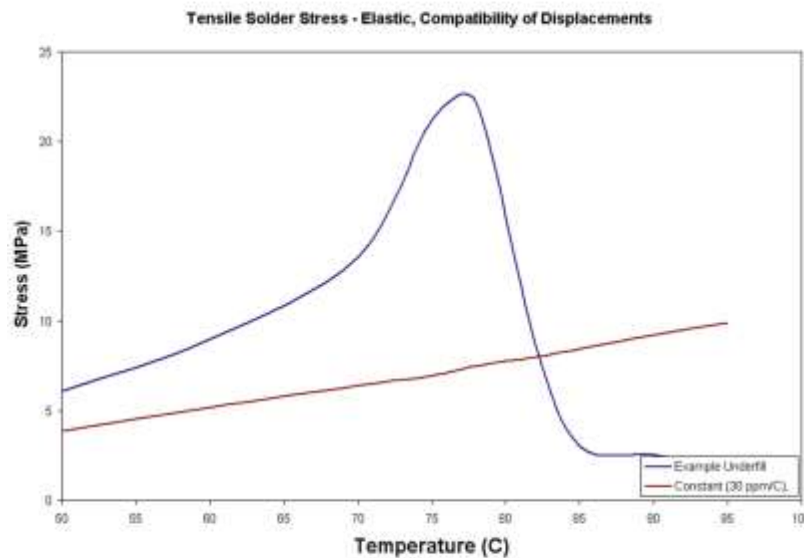
The other major challenge for identifying the root-cause of low Tg underfill failures was the random location of the failure site. Typically, in most electronic applications, thermo-mechanical fatigue of solder joints is due to shear stresses, which peak at the maximum distance from the neutral point (i.e., the corner). However, low Tg underfill failures, for reasons I will explain in the next paragraph, are driven by tensile stresses. And failures due to tensile stresses can literally occur in any flip chip solder bump. When the flip chip device has 1000 bumps, it can be like trying to find a needle in a haystack.

So, you might be thinking at this point, why tensile stresses? What tensile stresses? Tensile stresses arise in the solder bump when the operating temperature of the flip chip device goes through the glass transition temperature (Tg). An explanation of why can be seen in the next two charts.

As a quick refresher, the glass transition temperature marks the midpoint between a thermoset polymer being hard and brittle to soft and squishy. As underfill approaches the Tg, the CTE starts to increase dramatically (from 30ppm/C to 80ppm/C) with no corresponding decrease in modulus. This delay occurs because changes in the coefficient of thermal expansion (CTE) in polymers tend to be driven by changes in the free volume while changes in modulus tend to be driven by increases in translational and rotational movement of the polymer chains. Increases in CTE tend to initiate before decreases in modulus because lower levels of energy (temperature) are required to increase free volume compared to increases in movement along the polymer chains.



Once the underfill has a high CTE and a high modulus, increases in temperature will induce a high tensile stress since the stiff underfill is now expanding at a rate greater than the solder bump (see the next chart). However, as the temperature reaches the Tg, the modulus now begins to drop by orders of magnitude. This results in a very soft material that, even though the CTE is still high, does not have the stiffness to maintain stress on the solder bump. This behavior means that there is a very narrow window in which low Tg underfill would cause failure in electronic devices.



In fact, a change in operating temperature of just 5C can reduce (or increase) times to failure by up to 5X to 10X. This is partially due to the narrow window in which CTE and modulus is high and also partially due to the unique nature of tensile and compressive stresses on thermo-mechanical fatigue behavior. Previous research on solder has shown that when the mean stress changes from tensile to neutral (not tensile or compressive), the cycles to failure can change between 10X to 100X (this is when the stress range is the same for both conditions). And when the mean stress changes to compressive, time to failure can be almost infinite. And this infinite condition is what

happened when initial testing seemed to show that devices with low Tg underfill were bullet-proof.

So, what have we learned and where do we go from here? First, in regards to low Tg underfill, all tests based on JESD22-A104 are WORTHLESS. Let me repeat that: WORTHLESS. Second, the industry may soon reach a decision point. The immediate response to the low Tg underfill fiasco was to ban low Tg underfill from all products (assuming the OEM customers knew about it). The result is that there are relatively few companies still using low Tg underfill. Or, if they are, they are making a more concerted effort to ensure the operating temperature is far away from Tg. This has resulted in a number of mid-Tg products, where the Tg is not as low as the 70C that caused failures but not as high as the original 125-150C.

However, these higher Tg underfills cause problems not only for low-K dielectrics, but also in regards to warpage of the flip chip. The larger the silicon die, the larger the warpage. If the warpage reaches a critical limit, it cannot attach to a lid or some other thermal solution. The tipping point seems to be when the die size exceeds 25x25. Many ASIC designs and devices for telecom/enterprise applications are reaching this limit. Will they go back to low Tg blindly? Will they make an effort to understand low Tg and apply it correctly? Or will they go in a completely different direction (except TSV; that has been pushed out for another three years). Time will tell.