IWLPC Dinner Keynote

Dr. Thomas H. Di Stefano, Centipede Systems, 30:02 minutes, Oct. 15, 2008

Ten orders of magnitude of productivity gains in half a century. That has never happened in the whole course of human history. That drive is why we have this computer, that projector, your iPODs. All the thing that are essential to how we live today came from one simple fact: the integrated circuit and its productivity gains, year-after-year, following Moore's Law, which doubled the number of transistors every two years and that of course leads to a cost reduction in the cost of the device. That's the basis for our modern civilization.

Think about it. If you didn't have that, you would have to give up all your electronics, everything, the phones—maybe you would have a black phone—but that is the basis of how we live today. It's all based on productivity gains of the integrated circuit. Now the packaging folks have contributed somewhat. The productivity gains in packaging. Why doesn't packaging contribute more to the overall progress of the field? Well, one of the reasons is that packaging always had a 19th century feel to it—bending metal, stamping metal, smashing wires against things then join them hot. It has a one-at-a-time make-things feel to it, whereas the integrated circuit's all processing. The difference is processing versus one-at-a-time fabrication. The promise of wafer-level packaging is to break free of that constraint. It allows us to process some, a portion of, or all of a package by similar methods used to process integrated circuits instead of one-at-a-time fabrication. Now, in WLP, one of the things that has helped the field is the chip-scale package. Let's look back a bit at what has happened in packaging. There are a few main things that stand out: Every two decades, a new technology surfaces in packaging, driven by density—getting more contacts to the semiconductor device Through hole technology, then surface mount, then area array packaging. When we get to area array, chip-size packaging means that the package is under the shadow of the chip.

And for the first time, we can actually make the package on the wafer…previously the package was huge and the chip small, but here, chip-size, we can actually make the

package right on the wafer and process it and enjoy the benefits of the productivity gains of processing versus one-at-a-time, making wire bonds and what have you.

A deeper look at that, packaging advances have really come about by the need for density—more connections to the chip. It propagates through all levels of semiconductor interconnect. Through holes were used with DIP packages, with one lead going down each through hole. That was fine until the package got larger and larger and we couldn't get enough through holes on the board to support the number of I/Os on the chip. So guess what, the next generation said we don't need a through hole for every lead; let's just surface mount the leads and wire them out to vias. That worked fine until the number of leads we could get around the periphery of the chip, on say a QFP, maxed out at about 20 mil spacing for the leads. And that was it, you can't get more around the package. And that caused us to pop the leads into the inside of the package and wrap the leads right around the chip. In the early days, we called that fan-in because the leads that came in were underneath the chip. And in that generation we're pushing the substrate density to match the area array grid pitch, then it led to micro-via substrates. Then an important thing happened: For the first time, we can now process the package or some or part of the package on the wafer. Going through this, I'm going to make a distinction between wafer-level package processing and a complete package.

WLP is a paradigm for how you make packages. You package them by processing, hopefully on the wafer. If you end up with a package that's larger than the chip, so be it, but you've made use of the fact that a portion of the package is fabricated on the wafer and that's the wafer-level paradigm. That allows all the benefits that accrue to the IC fabrication that enjoyed learning curve advances year upon year. Process improvements have brought lower cost, higher density, and more functionality. The factors here are that in wafer level we're processing versus assembly. We're not wire bonding or making individual leads, we're doing the whole thing 10 billion devices at a time or 50,000 leads at a time, driven by cost reduction and hopefully we have a steeper learning curve than in our 19th century past of packaging.

Very importantly, these techniques are adaptable to a diverse set of packages, all the way from MEMS to integrated circuits to stacked chips. The same paradigm can be used for any packaging technology. Here's an important thing that actually I was wrong on initially I looked at wafer-level packaging as a package technology—that there was such a thing as a wafer-level package. There's not really; it's a process for making a package and that's a very important distinction: It's a process, not a package technology that can be applied to any package technology, so long as a portion or the entire package can be made on the wafer.

Today, WLP enjoys exceptional growth. Jan Vardaman at TechSearch International projects 14 percent growth in the sector. Jan, I don't know if you have to redo your numbers in light of last week or so, but it certainly stands out as one of the bright areas in electronics, because not only does it reduce cost - which is always important - but it gives a smaller sized package; it provides highest performance along with additional functionality. What has happened is that WLP has proliferated into a range of package types: MEMS, stacked chips, we'll take a look at a few, rather than going linearly in a path through larger-and-larger chips, DRAMS, processors. The techniques of using processing to get on a steep line curve apply to all packaging, as long as you can process it on a wafer.

This is an illustration from my previous company, Tessera, showing a camera chip on a wafer. The whole wafer is made at one time with a camera chip. There are additional parts added to it. I believe this lens is added later. The lens can actually be added at the wafer level. This, as opposed to a mechanical assembly, where there are lots of piece parts, a machine and this-and-that. It's obvious that by using processing we can get the cost of this down to unbelievable numbers—like a dollar for a camera. That's phenomenal! All this by using wafer-level techniques, and that's not the end of it. It's on a learning curve where the costs keep going down and we can add more functions into that. So MEMS of various types, cameras, pressure transducers, complex systems for chemical measurement can be made and packaged on a wafer.

Another area that has come up in wafer-level packaging that promises to be quite important is stacked chips. This shows a chart going back six years that Joe Fjelstad quite presciently looked at the next thing that's important is stacked chips, goodbye area density, and that's happening now.

I believe every memory company in the world is looking at stacked chips - through silicon via (TSV), other technologies to stack chips. I've used published work from one of the pioneers in the field at ALLVIA, showing an ALLVIA stacked set of chips. If you look at the chip in a schematic cross section you would see a stack of these through vias, through silicon, where the through via connects to the successive chip above it. The bottom has a slightly larger through via because that's the one that gets soldered down to the board.

This is an area that could really have explosive growth if it fulfills its promise to give us a higher density for memory. Then the rest of wafer-level packaging is more or less in the linear and rf device sector. I want to take you through that and then some of the limitations. These are CSPs called MicroSMTs from National, one of the pioneers in the field. They have put many of their analog and mixed-signal chips on wafer-level packages which is really not much more than redistribution of bumped wafers, limited to less than 3 mm because of the thermal expansion mismatch between the chips and the low-cost substrates they're mounted to.

More recently, Lee Smith from Amkor, has presented a roadmap for increasing the size of chips from about 10 square mm up to about 42 square mm, and with increasing pincount. I'm not sure of the reliability, that's something we'd want to ask Smith about, but certainly increasing size. Amkor's view of the market, which Lee graciously provided, is that this is a rapidly growing business, confirming Jan Vardaman's thesis. WLP is a bright spot - 1.3 billion shipped over the last several years; that's a serious volume of devices and these are full redistribution layer devices with solder bumps and pushing the grid pitch down to about 0.4 mm, quite a bit of advance in the small chip arena.

Let's look at the overall picture of where these devices fit. We're leaving behind now MEMS, which are their own special case, and stacked chips. We're really looking at WLP individual die. Most of the activity in wafer-level dies is in this little box: passives and mixed-signal devices, up to about 25 I/Os and 3 mm on a side, limited by reliability, because these chips are not under-filled and the I/Os are limited by the density on the substrate.

At 0.5 mm grid pitch for a micro-via substrate, we're really in that confined region of the market, and WLP has done very well there: cost reduction, small size, all the good things, but still confined to that space. To break out into the larger market, we need to find a way to make these packages so they don't need under-fill - break that under-fill barrier - and go to larger die sizes, DRAMs and Flash memory especially, and possibly, eventually, microprocessors.

Many chips are too large for WLP because of the thermal expansion mismatch between chip and substrate. The question here is how can this domain of WLP be expanded more rapidly? What are the limiting factors and what can we do about it? WLP has not really penetrated DRAM. When it does, it's mainstream packaging. And the reasons are that burn-in and test for these dies is expensive; it's not there yet. Handling the dies, testing is difficult and expensive. A reliable solder test technology for these large dies is not yet practical, although there are many candidates in the wings.

Over the last, 10 years memory suppliers have made a run at providing wafer-level packaged parts and have fallen off the track for two reasons: burn-in, test and handling, and a reliable solder joining technology. It's not there yet, and, of course, in the DRAM and Flash memory markets, cost considerations are extreme, so it's got to be no more expensive than current technology.

Let's look a little bit at the problems we're facing. Over the last 10 years, full wafer burn-in was assumed to be necessary for wafer-level packaging - certainly for memory parts, but it stalled. It's just not getting there fast enough, and it's not because of the lack

of effort over that decade. The problems are really pretty basic. If you have 50,000 contacts over a hot die with a hot substrate or a hot probe attached to it, the pads move relative to the probes, move quite a bit, several mils, four mils in a simple case, more in a lower-cost case.

The cost has to be less than $50,000—not the $150,000 or $200,000 that a 300 mm wafer probe card costs today. You can't just take a wafer probe and reduce the cost for this problem; there's just too much of a cost reduction. It's got to cost well less than $50,000 because of the numbers of these things involved. It's not for lack of trying; it's a very difficult problem, and it's one of the two problems limiting wafer level in memory.

We're seeing some activity, rather than probe on the wafer, to dice the wafer, throw out the bad dice, and mount the good ones in a tray for burn-in and test. It greatly simplifies the process and cuts costs accordingly. You're not testing bad dies, you're just testing good ones; you have a standardized form factor and standardized automated handling; and a much simpler contactor. By testing devices in a tray, we can break through the bottleneck of how you test wafer-level packaged parts.

Wafer-level packaging does not mean that we have to do everything on the wafer. You don't have to test the wafer. Just make use of parallel processing to get the cost down and functionality up. Benefit by the learning curve. How do you test it? If test-in-tray works, it's low cost, do that. Certainly, looking at the analogue or the analogous situation of standard parts, standard parts are tested individually; it has not hurt productivity gains over the years.

So this is an alternative to the assumption that a wafer-level package has to be tested with a full wafer burn-in. It doesn't! The second problem is a low-cost, reliable solder-test technology, and there are many that have been tried and the menagerie of possibilities is amusing. There is anything you can think of and some that are still in the lab are being tried to get a low cost and yet reliable solder attached. There are proven technologies that are highly reliable; they're just way too expensive. There's stuff we can do that's cheap,

it just doesn't work! Or it sort of works. There's nothing ideal that's really bullet proof, low cost, reliable and applicable to memory parts. It's not quite there yet; it's close but not quite there. There's a lot of work going on, but you don't see DRAMs in high volume with wafer-level packages.

One I will go into I know a little bit about: This is the Tessera wafer-level package. The interesting thing here is not that Tessera had a wafer-level package—this is a five-inch wafer—100 percent yield, it's highly reliable, full wafer, pound down the plains, and you can put additional wiring plains in it. It's been available for a decade on a shelf! It's too expensive, too expensive to make it into the marketplace for DRAM or a high-volume product. Look at the wafer-level device. The device is fabricated on a wafer. The whole wafer at one time, the little flexible link made by injection molding, like expanded metal, that link, to get these little, flexible leads. This is an x-ray shot of those leads. It shows the interior of the wafer-level package.

My expectation when I was involved was that it would find a use in the market and it hasn't. Why not? Cost! Wafer-level packaging has to cost less than the equivalent package. You can provide all kinds of performance and functionality, but still it has to be less expensive than what's available.

Another approach that is being looked at, but is not being pushed more strongly, is low CTE boards. Boards in Japan are getting the CTE down so the thermal expansion mismatch is quite a bit less than on copper-based boards. Invar-based boards are more expensive but certainly could reduce the CTE. Here a direction to watch is that it's possible that the problem can be solved by simply using low CTE substrates and just get it over with. No sign it's going to happen, but it could. If that happens, memory could go wafer level—at least from a reliability point of view.

The future: It's frustrating because of the lack of progress. Wafer level has promise, not only for cost reduction to get packaging on a learning curve similar to what ICs have enjoyed for decades, but beyond that. If we process the package, we can actually put

more things in the package. We can put power and ground distribution in the package. We can wrap global routes into the package. We can integrate more things in the package at a reasonable cost. One of the things we could do in package is to minimize the I/O explosion, really the power and ground explosion that we have in high-end processor chips.

If you look at a processor chip, they look impressive. The chip is a flip chip; it has thousands of I/Os - wow, that's really technology! - but, if you look at it more closely, almost all of those solder balls are simply power and ground, just to get power into the chip, because you can't distribute power easily and get it to the chip. If you strip those away, the signal I/Os go slowly over time up to about 500 and 1000 for the highest performance chips.

Something that I would look for a wafer-level package to do is to make use of the package to redistribute the power and ground within the package. Thick copper has a high expansion, but it's in the package, and cut down the number of power and ground contacts that need to be innate to the package. That can certainly be done, and the technology is available. I guess the question is, is it cost effective? I presume the answer is "not yet."

Another thing that comes up with designers is nonlinear RC time constant delays. And what's the play for packaging? What happens in nonlinear delays is that as the signal lines shrink down, both laterally and vertically, the resistance of the line goes up. It goes up inversely with of the cross sectional area of the wire. As the wire shrinks down by a factor of two, the resistance goes up a factor of four. Unfortunately, that scaling does not work the same way for capacitance. Capacitance has a logarithmic relationship to the diameter of the dimension of the wire, so the capacitance doesn't really go down significantly as you scale things down. That's a simple-minded version of it, but that's basically what happens; so as each generation, or each node of technology, shrinks the dimension of the wire further, the resistance of the wire increases and capacitance stays

about the same. The net result is it's harder and harder to propagate a signal through that resistive line.

Today at the 0.18 micron node, a signal can propagate 2 mm before the signal is dead, and that's it! Designers have elaborate schemes of local nodes, regeneration across a chip to march the signal across the chip. It causes delays and it causes complications. One possibility is to put long global routes up in the package. Certainly, clock distribution is a simple example. Put it up in the package and we have good, low resistance control of these nets in the package. The non-linear delay problem will scale with each node of the technology. It just gets worse until someone is going to have to do something about it, far from the design patches that people have applied because that's all they have. That's a natural for WLP. Put the global routes up in the package.

A view of the future, to put everything all together, is that you can have electronic modules or functional modules with chip-scale, wafer-level packaged parts, stacked chips, thermal management, optical interconnect all in a very small space because the packaged chips are chip size, stacked in the case of memory for density; substrates with micro-vias at 0.3 grid pitch and below, and substrates stacked, and you'll get enormous amounts of computing power, into a very small package. The chips could be stacked, or could be single chips, but they could very well be wafer-level packaged parts. The grid density is 0.3 and down, and some of the electronics subsumed up into the package. That's looking maybe far into the future, but that's what I see.

Let's step back from what we've gone through. I think the biggest conclusion that I draw from all this is that WLP is a pervasive paradigm; it's a process to get packaging, as much as possible, into parallel processing on the wafer. The fact that the package is not completed on the wafer doesn't take away from the benefits that accrue to that parallel processing: it's cost reduction on a learning curve. With WLP, we can add more functionality, more wiring, ground planes. We can do more with the package than simply connect the dots pad to terminal; we can actually make the package more a part of the integrated circuit. Coming back to today, for WLP to go mainstream with DRAM or

Flash, a few simple-minded problems need to be solved: cost-effective test and burn-in and handling and second high-density wiring boards that go with those parts need to be low cost. What I don't mention here is that of course it needs to be a reliable solder attach technology. With this, I hope I've given some insights into the overall WLP; hopefully, you've learned something. I know that in doing this, I did. Looking back over where it was and where the field is today. I think the field has more promise now than when I first looked at it. I looked at it as a package technology, linearly, for DRAM and processors, but it's broader than that: It's processing a package on the wafer by using a process rather than assembly to get cost, function and performance. Thanks you very much.