Transcript of 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 that 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 project here. 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, with a CSP we can actually make the

32 package right on the wafer and process it and enjoy the benefits of the productivity gains
33 of processing versus one-at-a-time, making wire bonds and what have you.
34
35 A deeper look at that, packaging advances have really come about by the need for
36 density—more connections to the chip. It propagates through all levels of the
37 interconnect. Through holes were used with DIP packages, with one lead going down
38 each through hole. That was fine until the package got larger and larger and we couldn't
39 get enough through holes on the board to support the number of I/Os on the chip. So
40 guess what, the next generation said we don't need a through hole for every lead; let's
41 just surface mount the leads and wire out to vias. That worked fine until the number of
42 leads we could get around the chip, around the periphery of the chip on say a QFP,
43 maxed out at about 20 mil spacing for the leads. And that was it, you can't get more
44 surface mount to the chip. And that caused us to pop the leads into the inside of the
45 package and wrap the leads right around the chip. In the early days, we called that fan-in
46 because the leads that came in were underneath the chip. And in that generation we're
47 pushing the substrate density to match the area array grid pitch, then it led to micro-via
48 substrates. Then an important thing happened: For the first time, we can now process the
49 package or some or part of the package on the wafer. Going through this, I'm going to
50 make a distinction between wafer-level package processing and a complete package.
51
52 WLP is a paradigm for how you make packages. You package them by processing,
53 hopefully on the wafer. If you end up with a package that's larger than the chip, so be it,
54 but you've made use of the fact that a portion of the package is fabricated on the wafer
55 and that's the wafer-level paradigm. That allows all the benefits that accrue to the IC
56 fabrication that enjoyed learning curve advances year upon year. Process improvements
57 have brought lower cost, higher density, and more functionality. The factors here are that
58 in wafer level we're processing versus assembly. We're not wire bonding or making
59 individual leads, we're doing the whole thing 10 billion devices at a time or 50,000 leads
60 at a time, driven by cost reduction and hopefully we have a steeper learning curve than in
61 our 19th century past of packaging.
62

63 Very importantly, these techniques are adaptable to a diverse set of packages, all the way
64 from MEMS to integrated circuits to stacked chips. The same paradigm can be used for
65 any packaging technology. Here's an important thing that actually I was wrong on
66 initially I looked at wafer-level packaging as a package technology—that there was such
67 a thing as a wafer-level package. There's not really; it's a process for making a package
68 and that's a very important distinction: It's a process, not a package technology that can
69 be applied to any package technology, so long as a portion or all of that package can be
70 made on the wafer.
71
72 Today, WLP enjoys exceptional growth. Jan Vardaman at TechSearch projects 14
73 percent growth. Jan, I don't know if you have to redo your numbers in light of last week
74 or so, but it certainly stands out as one of the bright areas in electronics, because not only
75 does it reduce cost—which is always important—but it gives a smaller sized package; it
76 provides highest performance along with additional functionality. What has happened is
77 that WLP has proliferated into a range of package types: MEMS, stacked chips, we'll
78 take a look at a few, rather than going linearly in a path through larger-and-larger chips,
79 DRAMS, processors. The techniques of using processing to get on a steep line curve
80 apply to all packaging, as long as you can process it on a wafer.
81
82 This is an illustration from my old company, Tessera, showing a camera chip on a wafer.
83 The whole wafer is made at one time with a camera chip. There are additional parts
84 added to it. I believe this lens is added later. The lens can actually be added at the wafer
85 level. This, as opposed to a mechanical assembly, where there are lots of piece parts, a
86 machine and this-and-that. It's obvious that by using processing we can get the cost of
87 this down to unbelievable numbers—like a dollar for a camera. That's phenomenal! By
88 using wafer-level processing and that's not the end of it. It's on a learning curve where
89 the costs keep going down and we can add more functions into that. So MEMS of
90 various types, cameras, pressure transducers, complex systems for chemical measurement
91 can be made and packaged on a wafer.
92

93 Another area that has come up in wafer-level packaging that promises to be quite
94 important is stacked chips. This shows a chart going back six years that Joe Fjelstad
95 quite presciently looked at the next thing that's important - stacked chips, and that's
96 happening now.
97
98 I believe every memory company in the world is looking at stacked chips—through via,
99 other technologies to stacked chips. They're not talking about it so I've got to use
100 published work from one of the pioneers in the field at ALLVIA, showing an ALLVIA
101 stacked set of chips. If you looked at the chip in a schematic cross section you would see
102 a stack of these through vias, through silicon, where the through via connects to the
103 successive chip above it. The bottom has a slightly larger through via because that's the
104 one that gets soldered down to the board.
105
106 This is an area that could really have explosive growth if it fulfills its promise to give us a
107 higher density for memory. Then the rest of wafer-level packaging is more or less linear.
108 I want to take you through that and then some of the limitations. These are CSPs called
109 MicroSMTs from National, one of the pioneers in the field. They have put many of their
110 analog and mixed-signal chips on wafer-level packages which is really not much more
111 than redistribution of bumped wafers, limited to less than 3 mm because of the thermal
112 expansion mismatch between the chips and the low-cost substrates they're mounted to.
113
114 More recently, Amkor, this is from Lee Smith at Amkor, has a roadmap for increasing the
115 size of chips, and these are real chips, up from about 10 square mm, up to about 42
116 square mm and increasing the pincount with reliability. I'm not sure of the reliability,
117 that's something we'd want to ask Smith about, but certainly increasing size. Now the
118 road map for that is the market. Amkor's view of the market, which Lee graciously
119 provided, from their view is that this is a rapidly growing business, confirming Jan
120 Vardaman's thesis. that WLP is a bright spot—1.3 billion over the last several years;
121 that's a serious volume of devices and these are full redistribution layer devices with
122 solder bumps and pushing the grid pitch down to about 0.4 mm, quite a bit of advance in
123 the small chip arena.

124 Let's look at the overall picture of where these devices fit. We're leaving behind now
125 MEMS, which are their own special case, and stacked chips. We're really looking at
126 WLP individual die. Most of the activity in wafer-level dies is in this little box: passives
127 and mixed-signal devices, up to about 25 I/Os and 3 mm on a side, limited by reliability,
128 because these chips are not underfilled and the I/Os are limited by the density on the
129 substrate.
130
131 At 0.5 mm grid pitch for a microvia substrate, we're really in that confined little box, and
132 WLP has done very well there: cost reduction, small size, all the good things, but still
133 confined to that box. To break out of that box we need to find a way to make these
134 packages so they don't need underfill—break that underfill barrier—and go to larger die
135 sizes, DRAMs and Flash memory especially, and possibly, eventually, the processors.
136
137 Many chips are too large for WLP because of the thermal expansion mismatch between
138 chip and substrate. The idea and the question here is why this domain of WLP expanded
139 more rapidly. What are the limiting factors and what can we do about it? WLP has not
140 really penetrated DRAM, which is mainstream. When it does, it's mainstream packaging.
141 And the reasons are that burn-in and test for these dies is expensive; it's not there yet.
142 Handling the dies, testing is difficult and expensive. A reliable solder test technology for
143 these large dies is not available.
144
145 Over the last, actually, 10 years memory suppliers have made a run at providing wafer-
146 level packaged parts and have fallen off the track for those two reasons: burn-in test and
147 handling and having a reliable solder test technology. It's not there yet, and, of course, in
148 the DRAM and Flash memory markets, cost considerations are extreme, so it's got to be
149 no more expensive than what's there now.
150
151 Let's look a little bit at the problems we're facing. Over the last 10 years, full wafer
152 burn-in was assumed to be necessary for wafer-level packaging—certainly for memory
153 parts, but it stalled. It's just not getting there fast enough, and it's not because of the lack
154 of effort over that decade. The problems are really pretty simple. If you have 50,000

155 contacts over a hot die with a hot substrate or a hot probe attached to it, the pads move
156 relative to the probes, move quite a bit, several mils, four mils in a simple case, more in a
157 lower-cost case.
158
159 The cost has to be less than $50,000—not the $150,000 or $200,000 that a 300 mm wafer
160 prober costs. You can't just take a wafer prober and scale it down for this problem;
161 there's just too much of a gap. It's got to cost well less than $50,000 because of the
162 numbers of these things involved. It's not for lack of trying; it's a very difficult problem,
163 and it's one of the two problems limiting wafer level in memory.
164
165 We're seeing some activity—unfortunately, no one's talking about it so I've got to do it
166 schematically—to rather than probe on the wafer, take the wafer, dice it, throw out the
167 bad dies, put the good ones in a carrier and test them in the carrier. It cuts your cost.
168 You're not testing bad dies, you're just testing good ones; you have a standardized form
169 factor, standardized handling, cut the cost of all that and you still need a contactor. This
170 is an area where several DRAM makers and a Flash makers are active. The technologies
171 are wafer level through via chips and it's a way to break through the bottleneck of how to
172 test wafer-level parts.
173
174 Wafer-level packaging does not mean that we have to do everything on the wafer. You
175 don't have to test the wafer. Just make use of parallel processing to get the cost down
176 and functionality up. Benefit by the learning curve. How do you test it? If test-in-tray
177 works, it's low cost, do that. Certainly, looking at the analogue or the analogous situation
178 of standard parts, standard parts are tested individually; it has not hurt their productivity
179 gains over the years.
180
181 So this is an alternative to the assumption that a wafer-level package has to be tested with
182 a full wafer burn-in. It doesn't! The second problem is a low-cost, reliable solder-test
183 technology, and there are many that have been tried and the menagerie of possibilities is
184 amusing. There is anything you can think of and some that are still in the lab are being
185 tried to get a low cost and yet reliable solder attached. There are proven technologies that

186 are highly reliable; they're just way too expensive. There's stuff we can do that's cheap,
187 it just doesn't work! Or it sort of works. There's nothing ideal that's really bullet proof,
188 low cost, reliable and applicable to memory parts. It's not quite there yet; it's close but
189 not quite there. There's a lot of work going on and, yes it's close, but you don't see
190 DRAMs yet with wafer-level packages.
191
192 One I will go into I know a little bit about: This is the Tessera wafer-level package. The
193 interesting thing here is not that Tessera had a wafer-level package—this is a five-inch
194 wafer—100 percent yield, it's highly reliable, full wafer, pound down the plains, you can
195 put additional wiring plains in it. It's been available for a decade on a shelf! It's too
196 expensive, too expensive to make it into the marketplace for DRAM or a high-volume
197 product. Look at the wafer-level device. The device is fabricated on a wafer. The whole
198 wafer at one time, the little flexible link made by injection molding, like expanded metal,
199 that link, to get these little, flexible leads. This is an x-ray shot of those leads. It shows
200 the interior of the wafer-level package.
201
202 My expectation when I was involved with this, was that it would find a use in the market
203 and it hasn't. Why not? Cost! Wafer-level packaging has to cost less than the equivalent
204 package. You can provide all kinds of performance and functionality, but still it has to be
205 less expensive than what's available.
206
207 Another approach that is being looked at, but is not being pushed more strongly, is low
208 CTE boards. Boards in Japan are getting the CTE down so the thermal expansion
209 mismatch is quite a bit less than on copper-based boards. Invar-based boards are more
210 expensive but certainly could reduce the CTE. Here a direction to watch is that it's
211 possible that the problem can be solved by simply using low CTE substrates and just get
212 it over with. No sign it's going to happen, but it could. If that happens, memory could
213 go wafer level—at least from a reliability point of view.
214
215 The future: It's frustrating because of the lack of progress. Wafer level has promise, not
216 only for cost reduction to get packaging on a learning curve similar to what ICs have

217 enjoyed for decades, but beyond that. If we process the package, we can actually put
218 more things in the package. We can put power and ground. We can wrap global routes
219 in the package. We can integrate more things in the package at a reasonable cost. One of
220 the things we could do in package is to minimize the I/O explosion, really the power and
221 ground explosion that we have in high-end processor chips.
222
223 If you look at a processor chip, they look impressive. The chip is a flip chip; it has
224 thousands of I/Os—wow, that's really technology!—but, if you look at it more closely,
225 almost all of those solder balls are simply power and ground, just to get power into the
226 chip, because you can't distribute power easily and get it to the chip. If you strip those
227 away, the signal I/Os go slowly over time up to about 500 and 1000 for the highest
228 performance chips.
229
230 Something that I would look for a wafer-level package to do is to make use of the
231 package to redistribute the power and ground within the package. Thick copper has a
232 high expansion rate, but it's in the package and not on the chip where it would have an
233 expansion mismatch. The planes redistribute power and ground in the package, and cut
234 down greatly on the number of power and ground contacts on the package. That can
235 certainly be done the technology is available. I guess the question is, is it cost effective?
236 I presume the answer is "not yet."
237
238 Another thing that comes up with designers, something they hit more-and-more as a
239 roadblock, is RC time constant delays. And what's the play for packaging? Let's take a
240 look at it and see how the package plays in this. What happens is that as the signal lines
241 shrink down, both laterally and vertically, the resistance of the line goes up. It goes up
242 proportionally with the area of the cross section of the wire. As the wire shrinks down by
243 a factor of two, the resistance goes up a factor of four. Unfortunately, that scaling does
244 not work for capacitance. Capacitance has a logarithmic relationship to the diameter of
245 the dimension of the wire, so the capacitance doesn't really go down significantly as you
246 scale things down. That's a simple-minded version of it, but that's basically what
247 happens; so as each generation, or each node of technology, shrinks the dimension of the

248 wire further, the resistance of the wire increases and capacitance stays about the same.
249 The net result is it's harder and harder to propagate a signal through that resistive line.
250
251 Today at the 0.18 micron node, a signal can propagate 2 mm before it's dead, and that's it!
252 Designers have elaborate schemes of local nodes, regeneration across a chip to march the
253 signal across the chip. It causes delays and it causes complications. One possibility is for
254 the critical global routes; put those routes up in the package. Certainly clock distribution
255 is a simple one to get across a chip. Put it up in the package and we have good, low
256 resistance control of these nets in the package. This is a problem that will scale with each
257 node of the technology. It just gets worse until somebody is going to have to do
258 something about it, far from the design patches that people have applied because that's all
259 they have. That's a natural for WLP. Put the global routes up in the package.
260
261 A view of the future, to put everything all together, is that you can have electronic
262 modules or functional modules with chip-scale, wafer-level packaged parts, stacked chips,
263 thermal management, all in a very small space because the chips are chip size, stacked in
264 the case of memory for density; substrates with microvia at 0.3 grid pitch and below, and
265 substrates stacked, and you'll get enormous amounts of electronics, computing power,
266 into a very small package. The chips could be stacked, could be single chips, but they
267 could very well be wafer-level packaged parts. The grid density is 0.3 and down, and
268 some of the electronics subsumed up into the package. That's looking maybe too far into
269 the future, but that's what I see.
270
271 Let's step back from what we've gone through. I think my biggest conclusion that I draw
272 from all this is that WLP is a pervasive paradigm; it's a process to get packaging, as
273 much as possible, into parallel processing on the wafer. The fact that the package is not
274 completed on the wafer doesn't take away from the benefits that accrue to that parallel
275 processing: it's cost reduction on a learning curve. With WLP, we can add more
276 functionality, more wiring, power and ground planes…. We can do more with the
277 package than simply connect the dots pad to terminal; we can actually make the package
278 more a part of the integrated circuit. Coming back to today, for WLP to go mainstream

279 with DRAM or Flash a few simple-minded problems need to be solved: cost-effective
280 test and burn-in and handling and second high-density wiring boards that go with those
281 parts need to be low cost. What I don't mention here is that of course it needs to be
282 reliable, a reliable solder attached. With this I hope I've given some insights into the
283 overall WLP; hopefully, you've learned something. I know that in doing this, I did.
284 Looking back over where it was and where the field is today. I think the field has more
285 promise now than when I first looked at it. I looked at it as a package technology,
286 linearly, for DRAM and processors, but it's broader than that: It's processing a package
287 on the wafer by using a process rather than assembly to get cost, function and
288 performance. Thank you very much.
289
290
291
292
293
294
295
296
297
298