Wafer Probe Acquires a New Importance in Testing

Wafer probe was once considered a method for saving packaging costs of bad die. Today, it's a critical element in process control, yield management, yield enhancement and customer quality—as well as overall cost of test. Moreover, full testing at wafer sort (probe), followed by testing for assembly-related failures at the package level, may not be far away.

By Mark Allison,
Electroglas Inc., San Jose, Calif.
[electroglas.com]

This is an interior view of a highly automated wafer probing system.

The paradigm for test is changing rapidly. Originally, for example, testing at wafer probe was employed for two purposes: The first was to weed out the bad die and save on the cost of packaging faulty devices.

This is still a concern, but with the high yields achieved on most commodity devices, this savings no longer offsets the cost of the additional testing equipment required.

The second purpose of testing at wafer probe was to provide yield feedback to the wafer fab quickly. Prompt feedback to allow correction in the fab process today is even more time sensitive. With much of the assembly and final test occurring overseas, thousands of additional wafers could be processed in the time between completion of the fab process and testing of those devices at a remote facility.

Historically, the probe card and interface technology were not sufficient to allow either a high level of parallelism or the high speed signals needed for at-speed device testing.

For memory devices with long test times for executing the test patterns, high parallelism was required to achieve cost-effective testing. In the past, testing at wafer probe was limited to a level of parallelism one or even two generations behind package test, primarily due to pincount and pin-density limitations on the probe card.

The development of high parallelism handlers was faster in coming than the development of probe cards required to test the same number of devices.

However, handlers designed to handle a very large number of devices in parallel may cost over one-half million dollars. Wafer probe systems could be much more cost effective if the level of parallelism could be increased.

Catalyst for Change

The days of device manufacturers selling only DIP and/or TSOP packaged parts are gone. With the increased pressure to get technology into smaller and smaller footprints, the size of the package has become a critical issue. In addition, the signal speed through these packages can be a limitation to overall system performance.

The concept of selling die instead of package parts is driving the need to ship these die with the best quality possible, which means full functional testing at wafer probe.

Termed Known Good Die, or KGD, these die will either get: a) packaged by

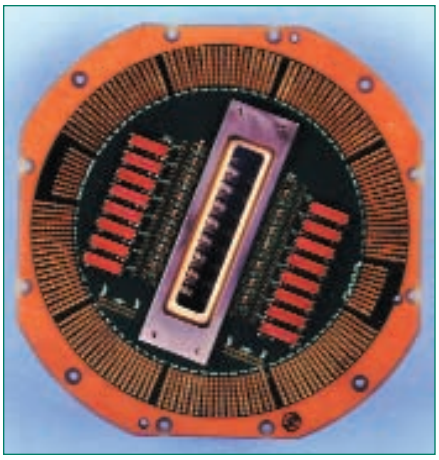

Figure 1. Probe cards for higher pin counts have to accommodate higher pad density where pad pitches are becoming finer and finer.

the end user in some type of custom package; b) mounted directly on a substrate; or, c) combined with other die in a multi-chip package (MCP).

With the emergence of MCPs, the yields of the individual chips in the package are combined to get the total MCP yield.

For example, if three chips are combined in a single package—with each chip tested at probe to a level that would have assured an 80 percent package test yield on the individual die—the total MCP yield would be 0.8 x 0.8 x 0.8 = 0.51 or a 51 percent MCP yield at package test.

Unacceptable Yields

This yield would be considered unacceptable and drives the need for full package-level testing at wafer probe.

If the die are tested to the same level as package test, the results are dramatically different. Assuming that the customer return level after package test is <0.01 percent (and it should be much less), then implementing this level of testing at wafer sort results in an MCP yield of >99.97 percent.

In addition to the end-user quality issue, it is also possible to optimize testing efficiencies at the bare die or wafer level. I/O compression and expansion are two examples. Either wide devices are tested as narrow parts to reduce the pin counts and tester cost; or narrow parts are tested as wider parts to potentially reduce the test time.

Built-In Self Test

In either case this is not always practical once the part is packaged. Additional test modes can include *built-in self-test* (BIST) that incorporates an engine on the chip, capable of exercising the device. Often this engine is accessed and controlled only through the use of additional pads that are only accessible before packaging.

Two new testing methods have emerged that reduce the need for package test to verify the device functionality at speed: One method sufficiently characterizes the effects of process variations on device performance in order to allow the manufacturer to ensure performance by monitoring and controling the process parameters.

The second method is to employ the BIST engine to exercise the device at

Figure 2. Wafer probers have become increasingly sophisticated and cost effective.

internal clock speeds. By using either of these methods for ensuring device performance, the device manufacturer can greatly reduce capital expenditures on expensive high speed testers and probe cards.

An option to address the concerns of MCPs and KGDs is the use of handlers for bare die testing. The complexity of handling bare, singulated die with sufficient accuracy to allow contact for testing has not been well addressed.

One approach has been to the use of bare die carriers. This approach has had marginal success to date. The additional handling needed to load/unload the carriers can represent a significant yield impact.

Pincount Increases

Pin count has increased exponentially on all devices, with the exception of memory. Increases in integration on-chip have fueled this growth in digital logic and mixed signal devices.

Total pincount in memory has increased at an even greater rate if we take into account the increase in parallelism used for testing efficiency.

Where once single device testing was dominant, now we see testing of 4-16 devices in parallel for digital logic and mixed signal, 64-128 devices in parallel for DRAM, 128-256 for Flash memory and even greater numbers for extremely small, low pin count devices.

Scrubbing Needed

These extremely high pincounts are a real challenge for probe card, prober and tester manufacturers. The increase in pincounts drives the need for probe technology that has sufficient force necessary to "scrub" through any non-conductive coatings on the pad, such as oxides.

The "scrubbing" action required to penetrate these coatings results in a force on the wafer that has a horizontal component as well as the vertical force. Vertical forces can range anywhere from 16g per pin for vertical (Cobra) type probes to 3g per pin for cantilever spring probes and under 3g per pin for lithographic probes.

The need to contact the wafer with thousands of pins is directly at odds with the newer technologies such as Al over Cu, Low-K and Active Circuits under Pads.

With thousands of pins, the force required to make a robust contact is in excess of 100kg. It is obvious that delicate handling of this contact is required.

This contact force is controlled by the wafer prober when it raises the wafer in the Z axis under the probe card.

To avoid damage to the devices, it is essential that the energy imparted to the wafer be minimized. This can be done using methods that enable the user to control the acceleration of the Z stage. The acceleration needs to be controlled in such a way as to reduce damage to the pads, low K dielectric and/or active circuits under pads, while at the same time optimizing throughput.

Probe cards for higher pin counts also have to deal with higher density of pads where pad pitches are becoming finer and finer, as shown in the table.

To achieve these finer pitches, pads are also getting smaller, which presents another challenge to probe cards and probers.

Pad Pitches Are Becoming Smaller

ITRS Roadmap 2004	2005	2007	2009	2011	2013	2015
Pad Size X x Y (µm)	35 x 65	30 x 55	25 x 45	25 x 45	20 x 35	15 x 25
% Scrub (Area/Depth)	25/50	25/50	20/40	20/40	20/40	20/40
Pad Pitch (µm)	40-100	40-100	30-80	30-60	30-60	30-60

Current memory probe yields are in the 80-95 percent range. Without redundancy, the yields would be more like 10-20 percent.

Due to their size, a high degree of accuracy is required to reliably hit the center of these small pads.

Motion Control

Probers have been developed that employ sophisticated motion control systems incorporating features including direct drive gantry designs with feed-forward and feedback controls, for a superior combination of accuracy, throughput, dynamic stability, and external disturbance rejection.

This level of attention to the design of the equipment is required to achieve the overall pin to pad alignment accuracies of better than ±1.5µm.

With more devices tested in parallel, the total probed area gets larger and larger—in some cases approaching one quarter of the wafer.

Maintaining probe card and probe tip planarity over an area this large, under these forces, is a challenge for both the probe card manufacturer and prober maker.

Planarity Essential

Stiffness of the probe card and probe tip planarity are essential; equally important is the mechanism for holding this probe card in the prober and the prober's ability to control the wafer, keeping it parallel to the probe tips.

Prober Z stages that are able to handle extremely high forces (up to 150kg) with very little deflection or tilt and the ability of the prober to accurately profile, and correct for, the unevenness in the wafer surface are very important.

This is particularly true for thinned wafers. Testing, prior to customer shipment, must be done after the wafers are thinned, since the device characteristics can change with thinning. For KGD sales this means testing thinned wafers at wafer sort.

Performance Characteristics

New probe card technologies are making rapid progress in terms of total pincount and performance characteristics.

Probe card companies are now entering the initial stages of development where they will be able, in the not too distant future, to offer the capability of contacting all die on the wafer at the same time.

While pin density on probe cards is increasing rapidly, signal bandwidth is making slower progress. While there are promising technologies emerging, most still fall short of accommodating the 1+GHz signals required for high-end devices. This may be an area requiring coverage using BIST, at least for the near future.

Conclusion

The evolution of testing, moving from package test to wafer test, is well underway.

Available probers, probe cards and testers have become increasingly sophisticated tools for maximizing throughput and yield thus reducing cost of test.

Mr. Allison is a 27-year veteran of memory testing. He is currently director of memory solutions for Electroglas Inc. He formerly served in various senior management roles at Advantest, Integrated Measurement Systems, Megatest, Micron Technology and Texas Instruments.
[mallison@electroglas.com]