Delta V Instruments

White Paper

A Comparison of Wafer Level Burn-in & Test Platforms for Device Qualification and Known Good Die (KGD) Production

John Darling

June 2003
Rev 1.0

Abstract

This paper defines the traditional package level burn-in & test (PLBT) technique and explains why a requirement for Known Good Die (KGD) is driving burn-in to the wafer level. After describing various wafer level burn-in & test (WLBT) platform solutions such as 1) Full Wafer Contact, 2) Wafer Multi-Probe, and 3) Sacrificial Metal Layer the paper compares the platforms and provides a summary analysis of current options.

Introduction

To date, the standard method of performing burn-in test on devices requires singulated die to be packaged and tested using Automated Test Equipment (ATE). The good devices are then placed into sockets mounted on custom designed burn-in boards. These burn-in sockets are designed specifically for high temperature applications. The loaded boards are then mounted into large chambers that control ambient temperature and provide a means for interfacing stimulus to the packages. At this point test vectors are used to stimulate the devices and a test routine is run for a number of hours as dictated in a qualification specification (MIL-STD-883-E or JESD22-A108-B). After burn-in the parts are unloaded and re-tested using ATE. Burn-in at the wafer level moves the process away from its traditional back-end location and allows burn-in data to be returned to the FAB process almost instantaneously. The cost savings in testing prior to singulation and packaging are enormous, not only monetarily but also in terms of time. The need for KGD is seen as a key stimulus in the development of WLBT platforms and has resulted in a wide range of possible solutions. This paper reviews the various platforms coming to the market and highlights the pertinent issues associated with each option.

Background

Historically, production device burn-in has been required on certain designs to eliminate early life failure (ELF) relating to the manufacturing process. Sample screening will detect process or design related defects while 100% screening will detect random die failures. Typically functional gross failures occur within the first 48 hours of stress testing with elevated ambient temperature of 125ºC and voltage levels 10% above nominal operating values. As failure data is analyzed, and manufacturing parameters adjusted, the production burn-in period can be reduced.

The ELF section of the graph shown in Fig.1 defines that region of the device life cycle that can be eliminated by performing 100% production burn-in.

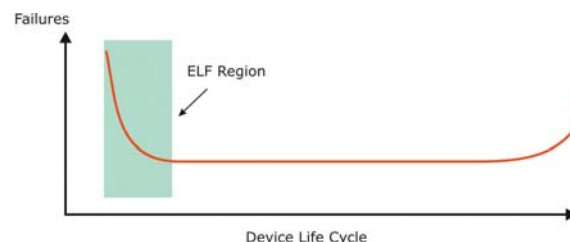

Fig. 1
Traditional Bathtub Curve
Defines Early Life Failure and End of Life Failure Probabilities

As the device architecture has increased in terms of gate complexity, while die geometry has been decreasing, it has become obvious that a package approach to performing ELF burn-in is no longer a viable solution. A test process flow which allows faulty die to be packaged only then to be discarded after burn-in results in a costly yield loss, especially if the failures could have been corrected. In addition to real failures associated with the manufacturing process, failures can also be introduced during the packaged testing. These types of failures are a result of socket reliability, operator intervention, spikes associated with bulk supply voltage rails, high power dissipation requirements and a failure to provide effective stimulus inputs to the device.

In addition, a growing demand for known good die (KGD) for use in system-on-chip (SOC) architecture, stacked memory applications and multi-chip modules (MCM) has defined a requirement for non-packaged burn-in solutions. This has lead to a consensus among the relevant engineering community that a move to wafer level burn-in is now an immediate requirement rather than an option to be considered. This shift in process will not be without several hurdles, the most significant being a contact method for interfacing the test electronics with the I/O pads and power planes at the individual die level.

An interim solution involves placing singulated die into custom-machined carriers, which are then placed into standard burn-in sockets. Traditional ELF burn-in with package level burn-in systems is then performed. While this is seen as an effective solution it has limited application due to the high costs involved in implementation. The use of carriers to perform burn-in also increases the risk of introducing false failures due to contact misalignment and handling errors. Scaling up to a large volume production application could be cost prohibitive.

The issue of interfacing test electronics to the device under test at the wafer level has spawned a plethora of potential solutions. All face stiff challenges, particularly relating to test capability, power dissipation, voltage rail tolerances, physical limitations (large quantity of die to be tested in a small working area), cost effective engineering, sustainable quality and correlation to ATE results.

Basic Platforms

Each approach to the issue of interfacing with the wafer must provide the necessary pin/pad assignment to perform sufficient test routines in order to qualify the bare die as KGD. There are three distinct approaches to achieving this requirement.

One method is to make contact with all relevant pads on each die. This facilitates testing at each die independently from all other die on the wafer. This type of interface is described as 'Full Contact'. This method has created most interest in the industry since the wafer under test would require no post FAB processing and by contacting all die on a wafer simultaneously, test time can be held to a minimum. A number of companies have based their WLBT platform designs on this approach.

The second method involves the use of modified probe hardware. A multi-head probe approach increases the throughput of die tested in parallel while controlling ambient temperature is achieved with traditional hot chucks. Companies basing their WLBT platforms on this method are primarily from within the probe manufacturing

industry. Since existing probe technology can be used as a base for the multi-probe configuration the initial cost to develop such a system is relatively low. However, having to step and repeat a number of times across a wafer increases the time to perform burn-in and is limiting in the type of testing which can be performed.

The third approach is to limit the number of interface connections to a manageable size and re-direct the interface design onto a sacrificial metal layer applied directly above the passivation layer on the wafer. This method requires post FAB processing to the wafer but does decrease the complexity of the interface to the test hardware. This approach is known as 'SML' and allows 100% die testing without 100% contact.

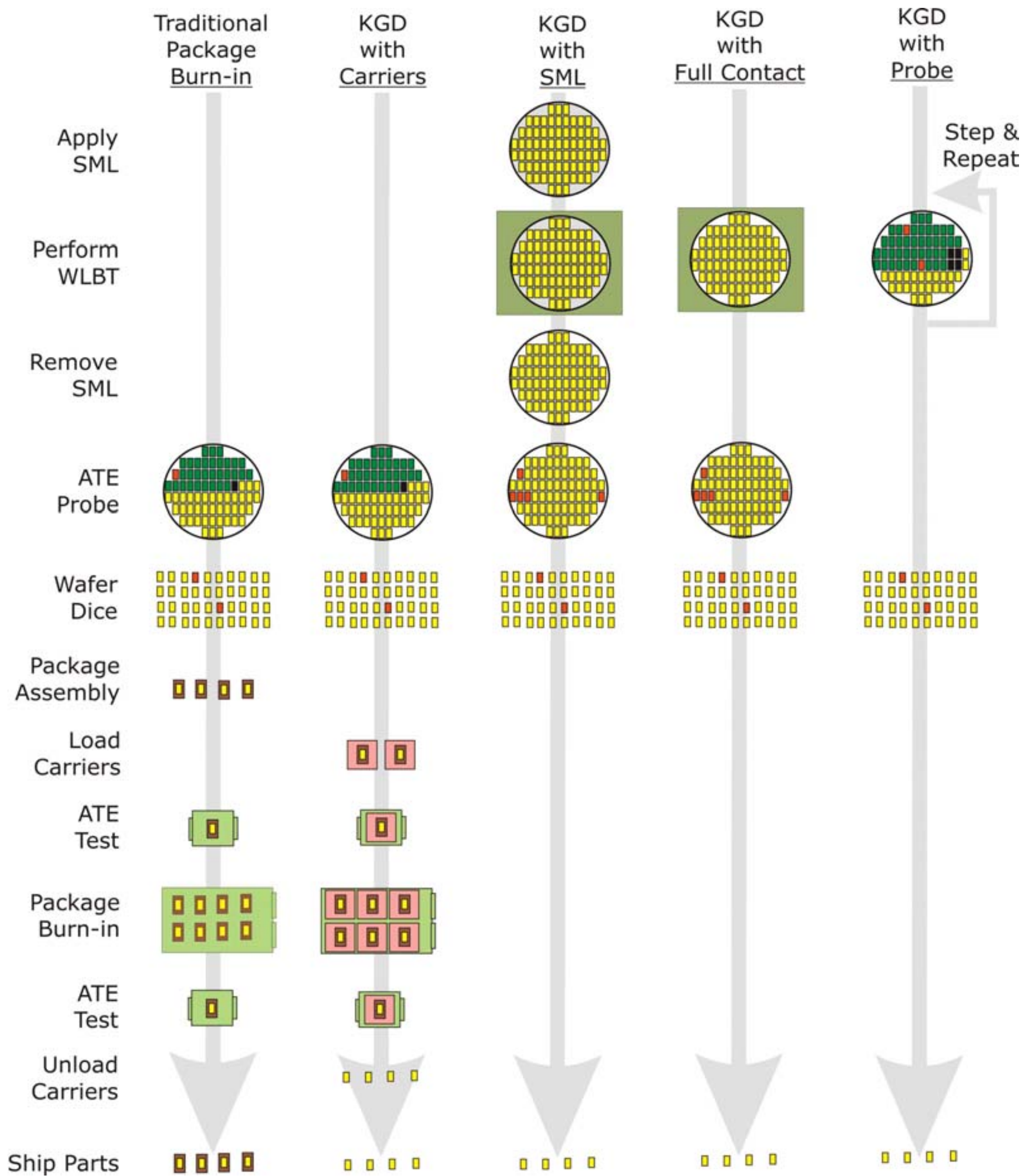

Fig. 2
Process Flow for each of the Test Platforms Described

Full Contact Platform

Full contact implies that all pads on each die will be connected via an interface to some form of test hardware. In reality only those pads required for test implementation need be contacted. This still results in a very large number of interface points. Since the test hardware is essentially the same for each full contact platform only the interface mechanics will be discussed at this time.

Interface mechanics rely on some type of interposer to provide the interconnection between the wafer and the test electronics. A full wafer interposer must be capable of:

- Providing a good planarity across an 8" or 12" diameter wafer – since the interposer will be positioned between the wafer and the test fixture all variations in the z-axis alignment must be compensated for as pressure is applied to both top and bottom faces of the interposer.
- Provide sufficient contact surface to ensure minimum resistance – the interposer must be capable of routing signal and power rails onto the wafer-under-test while making contact with multiple bond pads.
- Overcome alignment issues between the wafer and the test fixture – this alignment concern will vary depending on the type of interposer used. The test fixture and the wafer need to be precisely aligned to one another in order to match bond pads to test electronics. In some designs the interposer is physically attached to either the wafer or the test fixture while other designs have a interposer which can be sandwiched between the wafer and test fixture with no regard to the X-Y alignment.
- Exhibit a coefficient of expansion similar to that of silicon – since the wafer will be tested at ambient conditions approaching 200ºC the expansion across an 8" diameter wafer could result in a realignment of bond pads by as much as 50µm.
- Require relatively low contact pressure – as the number of interconnects rise within a given area so to will the force exerted in the z-axis. It is estimated that a force greater than 500lbs would be required to ensure reliable contact using micro-springs as an interconnect method.
- Be reusable, repairable and inexpensive – any interposer solution must exhibit an ability to be re-usable since ELF burn-in on production parts may require thousands of lot cycles.

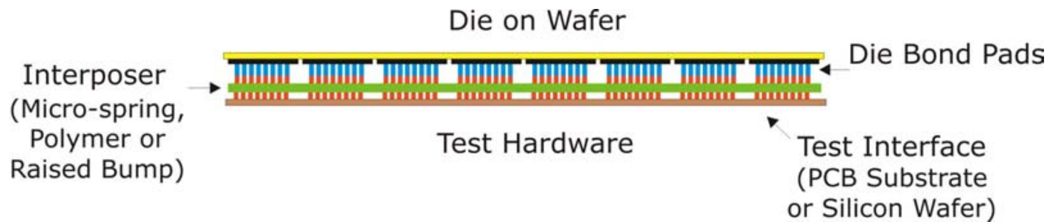

Fig. 3
Full Contact Platform uses an interposer medium for interconnect

One full contact platform relies on a miniature spring contact mounted between the wafer and the test hardware interface. These three elements must be held in perfect alignment using a cartridge type fixture that requires alignment equipment and a vacuum to achieve good contact.

A second full contact platform makes use of a micro-spring formed directly onto the die pads. An interface wafer, or PCB substrate, is then used as a pass through medium to the test hardware. The interface wafer is positioned directly over the test wafer and held in place under pressure. After testing has been completed the die are singulated and the micro-springs may then be used as an interface for packaging or connection to other die in a stack configuration.

A third full contact platform uses a custom designed interface printed circuit board (PCB) which interfaces to the wafer via an interposer material made up of thousands of conductive cylindrical pins mounted into a polymer. Since the cylinder diameter is smaller than the spacing between the device bond pads, alignment of the interposer is not an issue. Only the PCB and the wafer must be precisely aligned. This type of interposer material is consumable (must be replaced after 4-6 compressions) and expensive.

Probe Platform

Working from an established design base many probe manufacturers are expanding the number of test heads available to allow a multiple test approach. The step and repeat approach is a laborious method and one that is more in line with sample qualification than production burn-in. A high number of probe pins are contained in a small working area and make multiple touchdowns across the wafer surface. Reliability of the probe pins will be a major factor in the results from this approach. From the positive side, there is limited NRE involved in developing the test head and existing probe stations can be modified to handle the increase in test head capacity.

Interfacing to some type of test electronics has not been clearly defined, the interim solution being to link a modified prober to a piece of ATE equipment and measure IddQ levels. IddQ is the electrical current drawn by a device between clock cycles (inactive state). As die geometry shrinks the variation between IddQ measurements of a passing device and those of a failing device is also shrinking and may reach a point where IddQ is no longer a valid test.

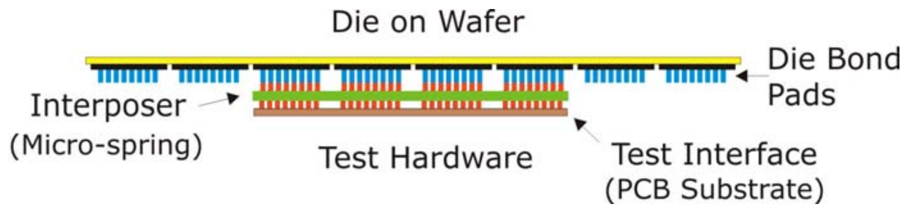

Fig. 4
Traditional Probing is expanded to handle Multiple Die

SML Platform

This approach to wafer level burn-in essentially copies the traditional package level approach of placing a number of devices onto a common platform and applying, in parallel, some stimulus. In the case of the die on a wafer, linking together a number of die forms several clusters. Each cluster is then fed, in parallel, the stimulus required for test. The linking of the die within a cluster is achieved using sacrificial metal layers, which link common die pads together and routes them to large geometry (2,000µm) touch down pads. The SML design also incorporates die isolation circuitry to limit yield loss in the event a die fails within a cluster. When a wafer is mounted into a test fixture, interface to the test electronics is achieved through a number of pogo contacts. After burn-in tests have been completed the sacrificial metal is removed using an etch process.

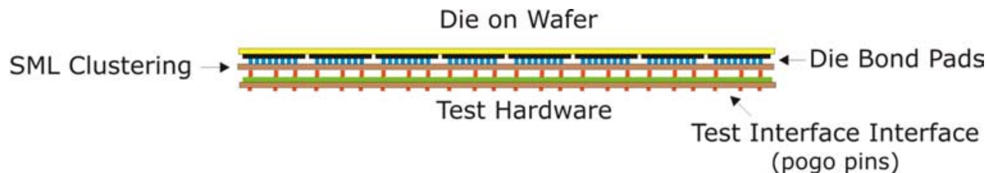

Fig. 5
Applying an Addition Process step to the Wafer FAB provides a Simpler Interconnect

Comparison of Benefits and Drawbacks Associated with each Platform

When selecting a burn-in platform for a specific application it is important to consider a number of factors, not all of which relate directly to the mechanics of the platform being considered.

- The level of testing required for KGD production – simple JTAG input or complex vector format?
- The level of testing required to qualify a product code – which tests can be ported from the ATE to run during the burn-in period, resulting in a saving in ATE cost?
- The monthly volumes associated with the application – with small volume runs it may be more cost effective to consider testing singulated die with the carrier approach.

- The Sustainable Life Period for the application – since much of the cost of full contact and SML is incurred up-front; a long life span (high volume) for the device being tested will provide a low amortized cost per die.

- Power Consumption – Die geometry and operating voltage levels are decreasing while complexity of die circuitry is increasing. These advancements require an increase in available power along with a method for removing the heat generated during test. Couple this with the requirement to test hundreds of die in parallel within the constraints of an 8" diameter wafer and it's clear not all test platforms will perform effectively.

- Available Hardware – Many of the full contact platforms are still in the development stage, the main hurdle is trying to find a suitable interposer method capable of making contact with the large pin counts while keeping the cost at an acceptable level.

- Consumable Test Fixture – If multiple product are to be tested at the wafer level the requirement to develop a new test fixture for each product may be cost prohibitive. The polymer-based interposer is a recurring cost, which may not be sustainable on high volume, low cost die. The design NRE associated with the SML platform will require amortization of a large volume device and may not be cost effective for small production runs.

The following table highlights the strengths and weaknesses of each platform.

	Full Contact	**Multi-Probe**	**SML**
Platform requires a custom test fixture?	Yes, each product code will require a unique test fixture design since the bond pad locations are not constant.	Yes, each product code will require a unique test fixture design.	No, a universal test fixture holding 2,800 interface pogo pins should accommodate a product family.
Test fixture needs to be mounted in a self-contained package?	Yes, due to the high interface count and the die bond pad geometry involved, it is necessary to seal the test fixture into a pressurized container to ensure precise alignment and good contact.	No, the test fixture is mounted directly onto the prober hardware.	No, due to the limited interface connections the wafer is aligned with a simple mechanical registration and pressure is applied by hand to lock down the interface.
Volume throughput?	High, since contact is being made with all die on the wafer test time is minimal.	Low, due to the limited number of die that can be probed at any one time. The step and repeat action will increase overall test time.	High, even though a limited number of contacts are used, the design architecture allows all die to be tested at the same time keeping test time at a minimal.

	Full Contact	Multi-Probe	SML
Interconnect Reliability?	Problematic, since the requirement to make contact with all active bond pads will demand an overall contact pressure of >500lbs for an 8" diameter wafer. In addition, the planarity issues associated with an 8" diameter silicon substrate cannot be readily compensated for using interposer materials such as GoreMateII.	Poor, the number of repeated touchdowns required with the prober method will reduce the reliability of the contact medium.	Good, due in part to the limited number of contact points being used and the dimensions of the pogo heads and the large touchdown pads on the SML. In addition the pogo pin movement in the z-axis will allow compensation for planarity issues associated with the 8" silicon substrate.
Requires additional wafer processing?	No, since all active bond pads are being contacted directly with an interposer material, no modification to the wafer structure is required.	No, essentially the probe approach is tried and tested and should not adversely effect the finished wafer. Probing does however tend to cause degradation of the bond pads and only a limited number of touchdowns are feasible.	Yes, the SML added to link the die into clusters and to provide large interface pads for the pogo heads must be removed. This is achieved using an etch process.
NRE Cost for Test Fixture?	High, since a large number of I/O lines must be routed along with power rails.	Low, due to the limited number of die being interfaced.	Low, since the fixture is universal There is NRE cost associated with the SML design but this can be amortized over the product life.
Contact Pressure?	High, estimated at 500lbs for full contact (30,000 bond pads) of an 8" diameter wafer.	Low, given that only a limited number of die are being contacted.	Manageable, since the universal test fixture limits the interconnects to 2,800 pogo pins. Overall force to contact 8" diameter wafer is estimated at 125lbs.

	Full Contact	**Multi-Probe**	**SML**
Scalability from 8” (200mm) to 12” (300mm)?	Not without a major re-think in the design concept. It is estimated that a 12” diameter wafer would require some 85,000 bond pad contacts resulting in an overall contact force approaching 1,200lbs. In addition, trying to maintain alignment of a 12” diameter wafer coupled with increased silicon expansion will become critical.	Time to test will become an issue since an increase in the number of step and repeat cycle will be necessary.	Yes, since the mechanics associated with the pogo contacts will remain unchanged. Moving to a 12” diameter wafer fixture will increase the overall contact force to an estimated 285lbs, with an increase in contact pogo pins to 6,500. Since the SML design is done with relatively large geometry rules scalability should not be an issue.
Cost of Ownership?	High, due to custom test fixtures, interposer material, which is expensive and consumable, and a need to provide a wide stimulus input. Ongoing costs include new test fixtures for each product code.	Low, in terms of capital equipment to perform the probing and the relatively low cost of design for the test fixture. The real cost is in the need to provide ATE for stimulus and the speed at which a wafer can be tested.	High, initial cost in capital equipment including universal test fixtures and SML design. An ongoing cost is associated with applying/removing SML to each wafer. The savings are gained when a complete product family can be configured to run on the same test fixture. With up-front knowledge of product family design, the SML can be modified for each product code rather than fresh design activity taking place.

Table 1
Comparison of Platforms

Summary

The concept of performing burn-in test while contacting all relevant die I/O and achieving near 100% coverage would be the ideal solution. The full contact platform promises this level of test but falls short due to the mechanical constraints associated with the die geometry and the magnitudes of power input and heat dissipation required to stimulate all die simultaneously. If these issues can be overcome at the 8" diameter wafer, they will only reappear when moving to a 12" diameter wafer.

The option to use existing hardware as in the case of a multi-probe approach delivers a tried and tested format. With some creative design techniques it may be possible to keep increasing the head count but at some point the multi-probe approach will encounter the same issues being uncovered by full contact platforms. In addition, this approach requires the use of ATE to provide stimulus that would be cost prohibitive for KGD production applications.

The use of sacrificial metal to form an interconnect architecture on the wafer allows a much simpler contact arrangement. Carefully designed isolation and fusible link circuitry can overcome potential yield problems associated with clustering die together. The main drawback is the need to perform further wafer processing steps after the standard FAB activity. Since much of the SML circuitry will be placed within the scribe lanes, complex SML circuitry may require an increase in these lanes – this will lead to a trade off between wafer yield and cost of test. Scalability does not appear to be an issue.

Each of the three platforms described in this paper offer a variety of positive and negative traits. When deciding on a platform to cover multiple product codes and product families compromises will have to be made.

All three approaches offer benefits but they are also in their infancy and have many design issues to overcome. Driven by a need to keep test costs down, the solution will not only be from a technical standpoint but implementation costs will also have to be considered based on anticipated volumes.

Since burn-in at the wafer level will no longer be part of the back-end post processing, the equipment can be located either directly in the FAB or in a BUMP facility. This moves the responsibility for performing burn-in test from traditional packaging providers and test houses to foundries.

In addition to the actual physical process of performing burn-in and test at the wafer level we should not forget that new standards must be written for this approach to burn-in. Many of the constraints associated with package level burn-in, such as temperature limitations due to packaging, will no longer apply. With the ability to perform burn-in on thousands of die in a single production run, the statistical quantities previously mandated for package level burn-in will need to be revised.

How to Contact Us

Website: http://www.delta-v.com

General Information: sales@delta-v.com

Technical Information: sales@delta-v.com

USA
Delta V Instruments
1870 Firman Drive
Richardson
Texas 75081

© Copyright 2003, Delta V Instruments. All rights reserved

This document is for information use only and is subject to change without prior notice. Delta V Instruments assumes no responsibility for any errors that may appear in this document. No part of this document may be reproduced, transmitted, transcribed, stored in a retrievable manner or translated into any language or computer language, in any form or by any means, electronic, mechanical, magnetic, optical, chemical, manual or otherwise, without prior written consent of Delta V Instruments

Product names mentioned in this document may be trademarks or registered trademarks of their respective owners and are hereby acknowledged.