HETEROGENEOUS INTEGRATION ROADMAP

2019 Edition

Chapter 17: Test Technology

Section 12: Burn-In and Reliability Testing

http://eps.ieee.org/hir

We acknowledge with gratitude the use of material and figures in this Roadmap that are excerpted from original sources. Figures & tables should be re-used only with the permission of the original source.

IEEE ELECTRONICS PACKAGING SOCIETY · ELECTRON DEVICES SOCIETY · IEEE Photonics Society · ASME SETTING THE STANDARD · semi

Section 12: Burn-In and Reliability Testing

The objective of reliability solutions and burn-in is to eliminate latent defects in ICs that will cause early-life failures and screen them out before the product is shipped to the customer. Reliability screens are critical to achieving the low failure rates required by high-reliability applications, such as automobiles. The visibility of automotive applications is increasing dramatically, as the electronics in cars includes not only engine and brake control, but also communications, internet access, entertainment, GPS, collision avoidance, and many other nascent applications. Reliability requirements in other mobile and mass storage applications is also increasing in importance, as the number of ICs and their transistor density increases. Latent defects are typically removed by process improvements, design improvements and accelerated stress methods during the test process. Reliability solutions are an optimization of 1) reliability defect density (RDD), 2) learning, reliability screens, and test methods (RS&TM) applications, and 3) design for reliability (DFR). The goal of the reliability solution optimization is to meet the reliability specification and the needs of the end customer while providing the best value for the reliability dollar spent.

Burn-In and Reliability Testing Goals

In reliability circles, customer satisfaction is measured by the field failure rate or failures in time (FITs). The cost of reliability screening has two components: manufacturing operations costs and yield. As such, these two components of the reliability cost equation are the primary challenges facing every reliability solution provider. In turn, manufacturing operations costs are also driven by three fundamental components – burn in duration, BIB/socket cost and equipment sophistication. The industry is still searching for a means to accelerate latent defects outside of the traditional elevated voltage and temperature methods. It follows that much progress has been made in detection techniques, but acceleration remains all about applying elevated voltage and temperature.

The component of reliability cost reduction associated with yield is severely biased towards elimination of "overkill"/"false rejects," which in many ways are tied to derivatives of the power solution. However, the primary source of false rejects stems back to the stress methodology, through the modeling assumptions, and ultimately finds its root in escapes from the manufacturing stress process.

The majority of market applications are most concerned with the early life component of the failure rate. Most latent defects that escape acceleration will fail early in the product life. The best way to guarantee a part received stimulus – and therefore did not escape stress – is simply to measure the outputs during stress. Defining terms: measuring outputs is called in situ stress, while measuring no outputs is dynamic stress. Obviously, the escapes component is less for in situ, and hence the early-life failure rate is lower. As anticipated, however, this lower failure rate does not come without cost. In situ stress requires functionality and functional test at stress conditions. Measuring outputs during stress also introduces a component of yield loss. Due to process variation, some portion of the distribution does not have sufficient margin to function at stress voltages or temperatures; however, these same parts may operate fine at application conditions. Although these parts may contain no reliability defects, in situ stress will fail these perfectly functional parts – hence over-kill. Determining the proper test method and interpretation of the test results are key ingredients of a successful in-situ burn-in strategy to ensure that the latent defects are identified and overkill is minimized. These same parts with "marginal margin" are the target of advances in detection techniques mentioned earlier. Achieving reliability requires trade-offs. In most instances, performance and yield hang in the balance.

Reliability defect density learning rate is the most cost-effective means of achieving the reliability demands of the marketplace. In itself, it is the by-product of the fundamental core practice in achieving profitability in microelectronics: yield learning rate. Stress conditions are no longer dictated by "technology nominal" specs but by system application conditions. Technology's recent inability to meet marketplace performance demands at reasonable power has forced systems designers to increase system application conditions (voltage and temperature) to compensate. Shifts in array Vmin operating range, NBTI-driven performance margin, and gate oxide integrity (time-dependent dielectric breakdown (TDDB)) as a result of the application of stress conditions still remain largely unexplained. As such, they dictate compensatory actions and/or reliability failure rate modifications. Even the standard thinking of metal electromigration for C4 and BEOL wiring requires careful scrutiny when confronted with the radical currents and powers conjured up by stress conditions.

DFR also has three key components: 1) technology design, 2) chip design (logical and physical), and 3) system design. In each of the three, the DFR work must strive for defect tolerance. In the case of technology design, leakage-induced power mitigation maintains an edge in importance over defect tolerance. Regarding chip design and DFR, power mitigation and fault tolerance are at par in design priority. Redundant element analysis and power dissipation

analysis consume considerable design engineering horsepower. At the system level, defect tolerance exists in the forms of error detection/correction and redundant elements.

In the arena of reliability screens and test methods, the literature is rich with techniques and methodologies with champions and supporting/compelling/biased data. Debates vary, depending upon the technology generation, chip/circuit type, design style, performance target, reliability requirements, and defect type. As long as excessive voltage and temperature retain the throne of defect acceleration, RS&TM will challenge the best and brightest minds in power delivery and thermal solutions. One must be able to accelerate defects while avoiding destroying the device – which is a change in precedence. In years past, stress conditions or actions that invoked or even hinted upon wear-out were to be avoided. The adage in the past was "one must be able to accelerate defects while avoiding the onset of wear-out." However, this is becoming increasingly more difficult in the face of stretched system applications conditions: sub-10 nm oxides; NBTI; marginal margin (that is, array Vmin); hundreds of amps and Watts; miles of copper wire; and billions of interconnects.

RS&TM are best categorized by separating them into wafer applications and package (or module) applications, and then further segregation into detection and acceleration techniques. This tiered structure will help to dilute the perennial argument between test and reliability regarding whether a field return is a test escape or an early life reliability failure.

Regardless of operational process step (wafer or package), acceleration techniques invariably must deal with potent power implications simply because acceleration requires temperature and/or voltage far in excess of application conditions – and leakage varies exponentially with both. The same is not true for detection techniques. In many instances, detection techniques employ conditions that reduce leakage (that is, VLV (very low voltage) or VLT (very low temperature)), and in instances where detection requires application conditions that exacerbate leakage, those conditions typically do not approach the level of acceleration conditions.

Burn-In and Reliability Testing Requirements

Technical challenges for the burn-in process are driven by increasing device pin count, decreasing package pitch, increasing device functionality and operating frequencies, dramatically increasing leakage current, and eroding voltage/thermal acceleration. In addition to burn-in, several alternate techniques such as IDDQ, high voltage stress, and wafer mapping are being used to try to improve device reliability.

Burn-in system technology must continue to evolve with device technology. The minimum device core voltage continues to decrease. Scan requires very deep vectors for large memories, while high power requires individual device thermal and power management. The burn-in process (system/driver/burn-in board/socket) will be challenged to meet speeds of the newest technology devices without some form of internally generated clock. Devices without DFT are requiring increasing I/O. The growing need for KGD continues to drive efforts for wafer level burn-in, KGD carriers, or additional stress during probe. Without continued innovation by the burn-in system manufacturers in cooperation with the IC manufacturers, all these trends tend to increase the cost of burn-in systems and sockets.

Device power and signal requirements are driving burn-in boards toward higher board layer counts, smaller traces, less space for routing, more complex processes and materials, higher test costs, and board reliability issues. Tight pitch on future devices will require new cost-effective, innovative interfaces between the burn-in sockets and the burn-in boards.

Burn-in sockets are undergoing major design challenges, as they must accommodate increasing contact count, decreasing pitch, higher currents, and higher frequencies. At the same time, sockets are a key component of an overall thermal solution designed to prevent high-power devices from self-destructing. A major challenge for socket manufacturers is to maintain low costs and short lead times while providing the technology to meet these new demands. Horizontally actuated contact design will be displaced below 0.5 mm pitch ball grid array (BGA) by vertically actuated contacts as pin count increases and existing socket materials fall short of increased mechanical stress requirements. New designs and new materials will be required for higher current carrying capabilities. Socket design will need to accommodate looser packaging specs in areas such as warpage and package dimensions, while coping with increased package size, thinner/more fragile packages, and reduced/non-standard/mixed pitches. Contact design will need to provide greater strength without a loss of electrical/mechanical performance.

Approaches to burn-in include traditional unit-level burn-in, system-level burn-in, wafer-level burn-in, and strip/array burn-in (Figure 1). In certain applications, system-level burn-in complements or replaces traditional device-level burn-in, but this typically involves a significantly increased cost, since the burn-in system, socketing solution and burn-in time tend to increase. Wafer-level burn-in technology continues to be developed, but has made only limited inroads against traditional package-level burn-in. The challenge here is to use techniques such as scan/logic and memory BIST (MBIST) to improve the technical feasibility of wafer-level burn-in.

Table 1: Burn-in Requirements

Year of Production	2018	2019	2020	2021	2026	2030
Packaged Part Burn-in						
Clock input frequency (MHz)	400	400	400	400	400	400
Off-chip data frequency (MHz)	75	75	75	75	75	75
Power dissipation (W per DUT)	600	600	600	600	600	600
Power Supply Voltage Range (V)						
High-performance ASIC / microprocessor / graphics processor	0.4–2.5	0.4–2.5	0.4–2.5	0.4–2.5	0.4–2.5	0.4–2.5
Low-end microcontroller	0.5–10	0.5–10	0.5–10	0.5–10	0.5–10	0.5–10
Mixed-signal	0.5–1000	0.5–1000	0.5–1000	0.5–1000	0.5–1000	0.5–1000
Memory	0.5-12.5	0.5-12.5	0.5-12.5	0.5-12.5	0.5-12.5	0.5-12.5
Maximum Number of Signal I/O						
High-performance ASIC	384	384	384	384	384	384
High-performance microprocessor / graphics processor / mixed-signal	128	128	128	128	128	128
Commodity memory	72	72	72	72	72	72
Maximum Current (A)						
High-performance microprocessor	450	450	450	450	450	450
High-performance graphics processor	200	200	200	200	200	200
Mixed-signal	30	30	30	30	30	30
Memory	10	10	10	10	20	20
Vector memory depth (M vectors – DFT/BIST SOC *2)	256	256	256	256	256	256
Maximum burn-in temperature (ºC)	175±3	175±3	200±3	200±3	200±3	200±3
Burn-in Socket						
Pin count	3000	3000	3000	3000	3000	3000
Pitch (mm)	0.08	0.08	0.08	0.08	0.08	0.08
Power consumption (A/Pin)	6	6	6	6	6	6
Wafer Level Burn-In						
Maximum burn-in temperature (ºC)	175±3	175±3	200±3	200±3	200±3	200±3
Pad Layout – See Probe Table						
Power consumption (KW/wafer)						
Low-end microcontroller, DFT/BIST SOC *2)	30	30	30	30	30	30
Memory	5	5	5	5	8	9
Maximum number of Signal I/O (Commodity memory)	45	45	45	45	45	45

Manufacturable solutions exist, and are being optimized

Manufacturable solutions are known

Interim solutions are known

Manufacturable solutions are NOT known

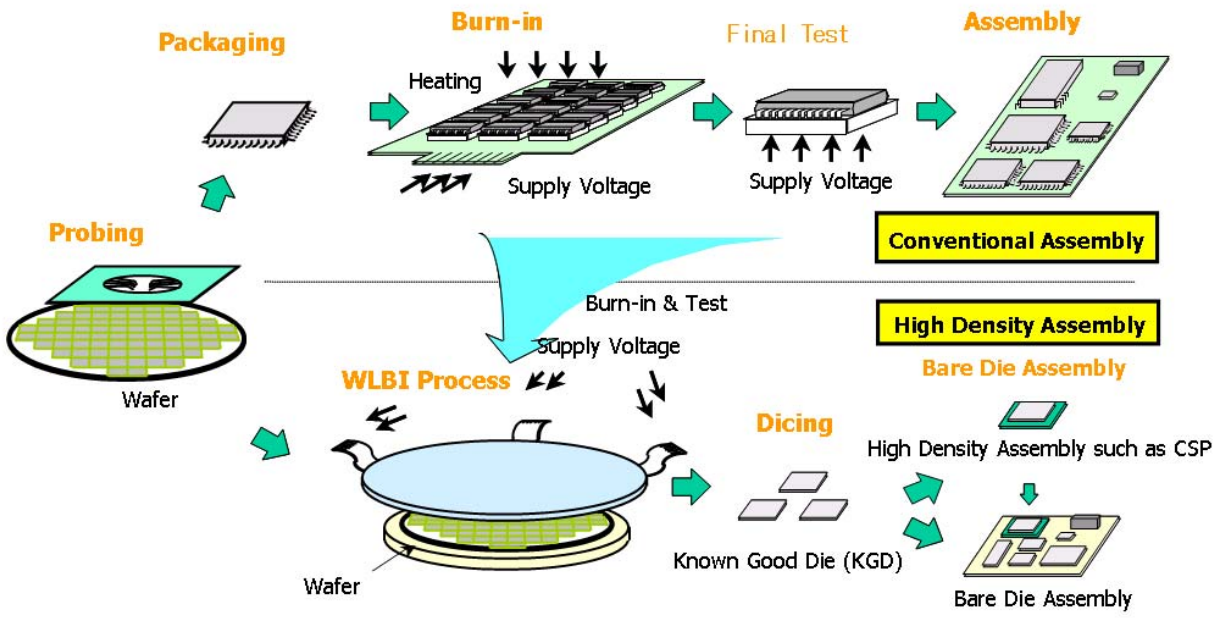

Figure 1: The Production Process with WLBI Compared with Package Burn-in.

Flash

The need for WLBI is increasing. The infant mortality rate is getting worse due to transistor scaling effects and new processing technology/materials for devices. Decreasing operating voltages and margins for devices are reducing the ability to use just voltage acceleration/voltage stress testing to guarantee reliability. KGD is becoming a more significant need by the customers due to requirements for chip-scale packaging and multi-chip modules, especially stacked die in mobile and mass storage applications. Reliability failures of the packaged part increase exponentially with the number of die in a multi-die package, so the need for reliable die before packaging is increasing substantially in importance. Decreased cycle time and the need for faster feedback of yield/defect information to the wafer fab can be assisted by moving burn-in earlier in the overall semiconductor process. Finally, detection and removal of defective devices prior to the packaging process eliminates packaging scrap costs based on intrinsic device defects.

There are two methods of performing a wafer-level burn-in. Some vendors use the term “burn-in” to refer to the application of a simple DC stress that applies opposite voltage potential to the internal nodes of a DRAM. This is typically referred to as wafer-level stress, as it applies a stress voltage for a short time, but does not apply the high temperature of burn-in. Actual WLBI requires full wafer contact and the application of high enough temperature over enough time to activate thermal defects, while also applying voltage stress with the device operating in “normal” mode. DFT functions such as scan or BIST are enablers for WLBI.

The challenge for DRAM for example, as a device well suited for WLBI, is to provide a burn-in environment for wafers that provides the same functionality, is as effective as package-level burn-in, and yet does not increase the cost of the final part. Leveraging the time spent in burn-in by using the burn-in environment as a massively parallel testing opportunity can effectively lower the overall cost-of-test.

Probing Technology for Wafer Level Burn-in

Full-wafer probing is a significant challenge, both technically and economically. The cost of a full-wafer probe tends to increase as the number of pads on the wafer increase and the pitch of the pads decreases. Contacting all the pads on a state-of-the-art wafer can require contacting in excess of 250,000 pads across a 300 mm wafer at a pitch of 60 microns or less over a wide temperature range. Intelligent use of DFT and pad placement rules by the semiconductor manufacturer can make this challenge less daunting. A WLBI micro pogo-pin contactor consists of a CTE-matched probe housing and pogo-pins with moving plungers at both sides. The pogo-pins stand vertically and have enough compliance and independent travel to accommodate height variations between adjacent contacts. Other vertical pin contactors operate in a similar manner. The probe pitch is technology dependent.

For a pitch less than 70 μm, MEMS technology by use of photolithography is an option. This technology, however, is very challenging for 300 mm wafers. While probing technology for tighter pitches is required, the intelligent use of DFT during pad layout may provide some relief by bypassing every other pad in order to double the probe pitch effectively, as compared to pad pitch. Application to high-pin-count and low-force probing due to low-κ materials will also be required. This will help drive new probing technology.

For contactor roadmaps, DRAM is selected as the target application due to its large predominance in general memory burn-in. DFT is considered for system LSI.

Other WLBI Technology Considerations

The current consumption of a wafer is increased by sub-threshold leakage from shorter transistor channel lengths and an increased number of transistors per unit area. The high temperature of burn-in also increases sub-threshold leakage. Therefore, the burn-in equipment must be capable of supplying over 1000 A of current per wafer in certain applications. Also, to manage current appropriately, wafer temperature control/uniformity becomes necessary. Finally, the burn-in equipment must be able to accommodate different quality distributions across each wafer.

BIST is capable of decreasing the number of pins under test per device, but die shrinks and tighter pad pitches can offset this advantage by increasing the total number of die and pads per wafer. The increased number of pins being tested also increases the force required to contact the wafer. In order to enable the use of WLBI through DFT functions such as scan, BIST, and JTAG1, the number of tested pins per device and total cost per device must be decreased and performance of the WLBI technology must be improved.

References

1. IEEE standard 1149, Boundary Scan