Highly Parallel Wafer-Level-Reliability Systems with Modular SMUs

Electronic Design

Jake Harnack

Mon, 2016-10-31 15:58

Reliability testing has long served as a method of ensuring that semiconductor devices maintain their desired performance over a given lifetime. As IC manufacturers continue to introduce new and innovative processes with decreasing device geometries, they need to ensure that the additional complexity from these changes doesn't affect the long-term reliability of their ICs. In addition, major technology trends in autonomous driving, cloud-based data storage, and life sciences are forcing IC suppliers to provide higher assurances of product reliability to their customers who work on mission-critical applications.

These two trends are driving semiconductor manufacturers to vastly increase the amount of reliability data they collect and analyze while decreasing the cost of test. When faced with this problem of more data at a lower cost, many reliability engineers find they cannot solve it using traditional reliability solutions. As a result, they're turning toward modular, flexible solutions that can scale to fit their needs *(Fig. 1)*.

Reliability Testing

Device reliability is typically modeled as failure rate over time, with the highest failure rates occurring immediately after manufacturing and again after the product has exceeded its useful lifetime *(Fig. 2)*.

Related

[Traditional Instruments: The Antique Collector's Dream](http://electronicdesign.com/print/test-measurement/highly-parallel-wafer...)

The left side of the graph shows early failures often caused by defects in the manufacturing process. These types of failures can be screened during production to minimize the number of defective parts sent to customers. However, the functional tests performed during production can neither identify defects that cause the device to prematurely wear out, nor offer insight into the product's usable lifetime. But reliability testing identifies these types of failure mechanisms and estimates the product's usable lifetime.

Reliability testing involves stressing a device at the extreme ends of the device's specifications—usually voltage and temperature—to accelerate device wearout and model the usable lifetime against known failure mechanisms. These tests can be performed on a wafer or packaged part. Wafer-level reliability (WLR) provides more data earlier in the manufacturing process without the cost and potential damage associated with cutting and packaging the IC.
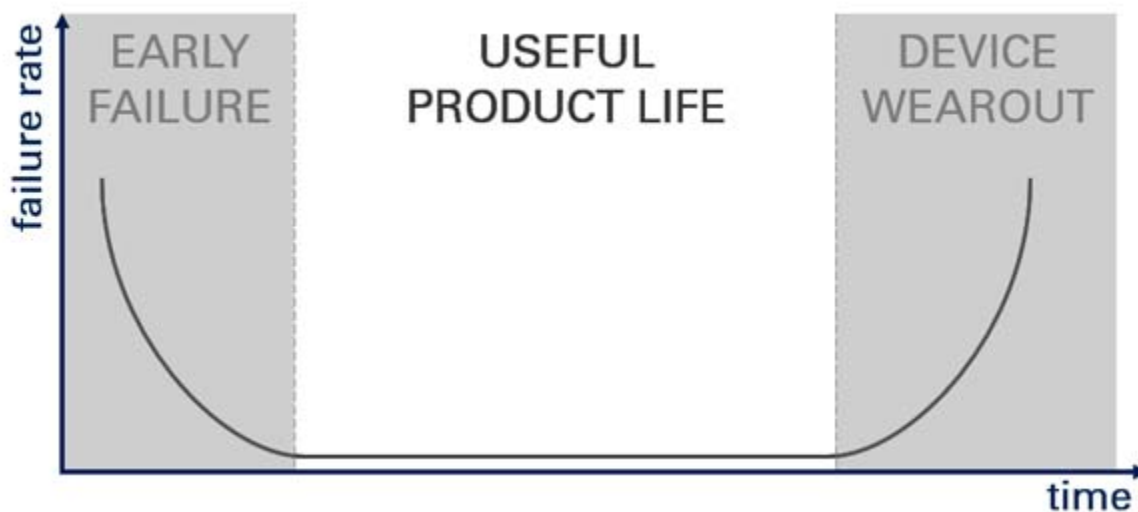
Wafer-Level Reliability

WLR is a type of parametric test that extracts information about the device's usable lifetime and long-term reliability. These tests typically aren't performed on the actual IC being developed, but rather a set of test structures or purpose-built dies that are built into the wafer specifically for gathering parametric data. These test structures consist of fundamental wafer elements such as transistors, capacitors, and resistors, which provide insight into the manufacturing process. Most WLR tests involve applying a stress, such as voltage or current, and measuring the response of the device to monitor for any signs of degradation. Common failure mechanisms include:

• Bias or negative bias temperature instability (BTI or NBTI)

• Hot-carrier injection (HCI)

• Time-dependent dielectric breakdown (TDDB)

• Electromigration (EM)

Traditional Approach to Building WLR Systems

WLR systems, which have been around for decades, vary in both measurement capability and architecture. Specialized WLR systems may involve high-frequency ac or pulsed stimulus. However, most CMOS devices are tested with dc instruments such as source measure units (SMUs), which supply the necessary stress and measurement capability for collecting parametric data. Historically, the two main approaches for building WLR systems involve either building a rack-and-stack system from traditional box instruments or buying a purpose-built turnkey system.

Rack-and-Stack Systems

SMUs are traditionally expensive, high-precision dc instruments that tend to limit the number of channels you can place in a standard test rack due to instrument size. Because of these constraints, SMUs are often combined with a low-leakage switching matrix to route signals from the SMU to dozens of test points while minimizing the noise, leakage current, and thermal EMF associated with relays. This approach works well when the serial testing of a small number of test structures generates statistically significant reliability data.

In addition, switching is a practical extension of a box instrument that historically has cost $5,000 to $10,000 per channel and been limited to 20 or 40 channels in a full 19-in. test rack. But, given the performance expectations for the relays, the switching subsystem is often a large and expensive piece of the WLR system.
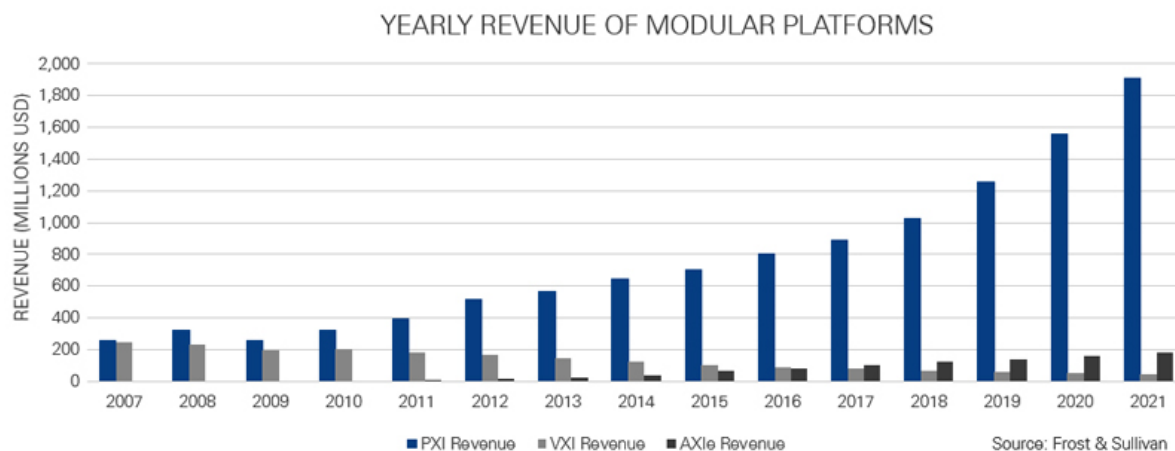
Turnkey Systems

The alternative approach is to purchase a purpose-built turnkey system that's prepackaged with all components, such as the oven, test rack, instrumentation, and software. Aligning your test requirements with the functionality of the equipment saves development and integration time, but requires a large capital budget.

These systems are often built with a fixed number of channels, hardware specifications, and software, and are serviced by the vendor. System vendors may sell separate systems for wafer and packaged reliability systems, or they may sell the same system for both applications regardless of the differences in test requirements.

Challenges of Traditional WLR Systems

The traditional WLR approaches of either buying purpose-built systems or building rack-and-stack systems from box instrumentation served their purpose for decades. However, many engineers are finding these architectures don't scale well to meet their new channel-density and cost requirements.


YEARLY REVENUE OF MODULAR PLATFORMS

Turnkey systems don't provide the flexibility needed to modify the test software or hardware as device requirements change, or the modifications are prohibitively expensive.

ck-and-stack systems are limited by the low-channel density of traditional box SMUs. This low density ates challenges for building high-channel-count systems with a small footprint, and often forces engineers to use a switched topology to multiplex the SMU to multiple pins. However, this switched topology quickly becomes a bottleneck because the pins are tested serially instead of in parallel. Therefore, implementing advanced stress algorithms that require constant stress and monitoring is impossible.

Because of these challenges, many companies have opted to build parallel test systems using modular instrumentation.

A New Approach for Building WLR Systems

The market for test instrumentation has changed dramatically over the past decade with the rise of modular platforms such as PXI *(Fig. 3)*. Modular platforms have grown increasingly desirable for building automated test systems because of their extensive I/O capability, compact form factor, and flexible software.
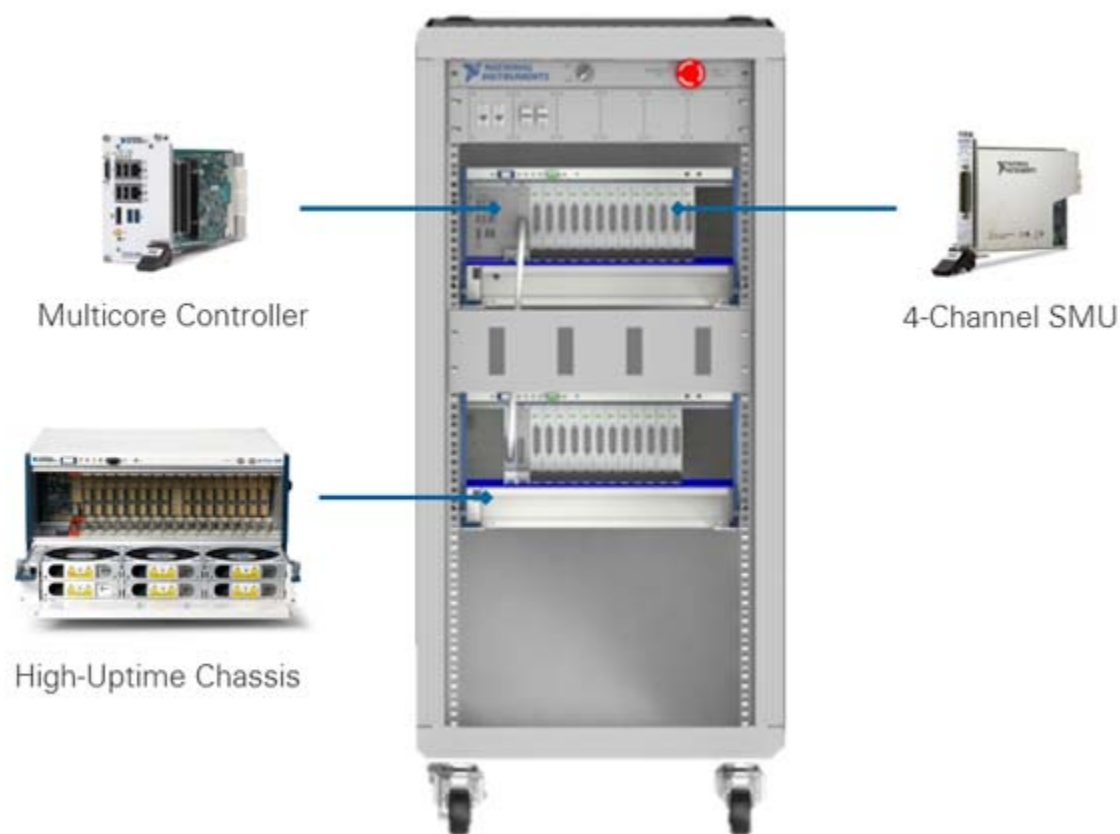
Using a modular approach, you can dramatically reduce the footprint of WLR systems without sacrificing measurement quality *(Fig. 4)*. The open software architecture allows you to define the functionality of your system, modify tests, and add hardware as your requirements change. This includes integrating the latest multicore processors, maximizing system uptime through health and monitoring tools, and adding I/O.

High-Density Source Measure Units

Modular platforms, such as PXI, allow you to build systems with hundreds of SMU channels while maintaining a reasonable footprint and cost per channel. With the high-channel density of these instruments, you can avoid routing your signals through a switching subsystem and instead connect each test pad directly to a high-precision SMU.
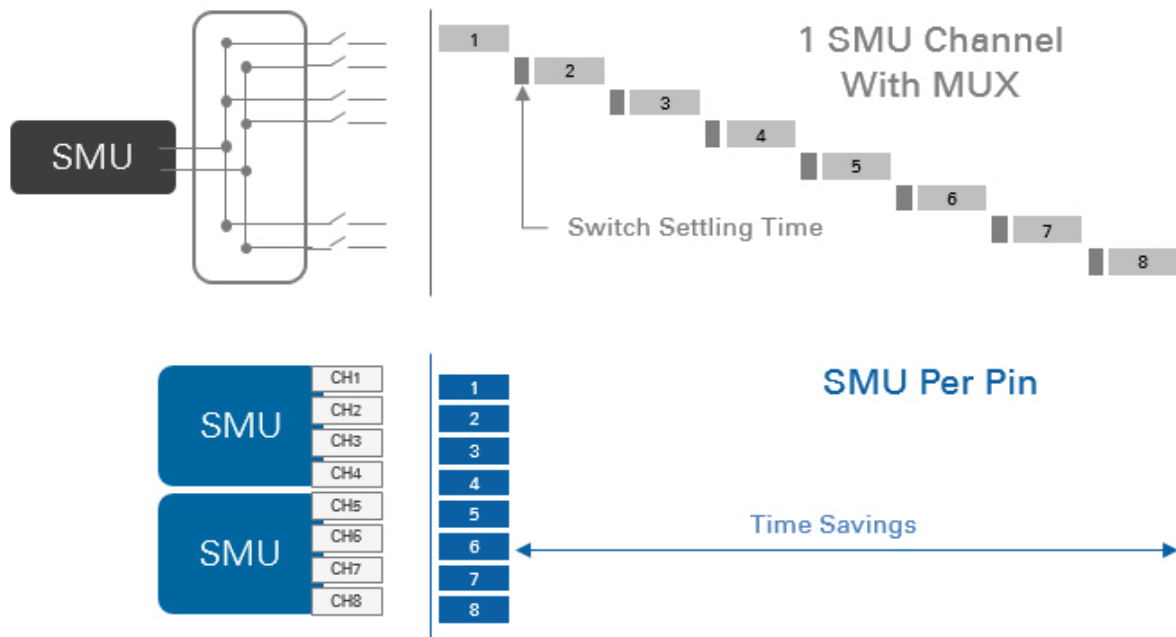
This "SMU-per-pin" architecture prevents the negative impact that switches have on signal integrity and test time *(Fig. 5)*. It also provides the flexibility to implement advanced stress-measure algorithms.

High Uptime and Serviceability

Ensuring system uptime is critical for both inline and offline reliability systems. If an inline system fails, wafer production can come to a halt. Offline reliability tests don't directly influence wafer production, but they do involve experiments that can run for several months. Ensuring a tester stays active and continues to acquire data is essential for the experiment's success, because a failed tester can lead to a failed experiment.

High-uptime applications often require systems with built-in redundancy for high-risk parts such as fans and power supplies. Building a test system with redundant, hot-swappable fans and power supplies allows you to mitigate the failure risk associated with these parts and ensures that the test system continues running after a component failure *(Fig. 6)*.

If the component is also hot-swappable, you can service the system without powering down the chassis and aborting the experiment. Furthermore, you can remotely monitor the health of your system for fan speed, temperature, power consumption, and other key parameters that may indicate an upcoming failure. By implementing these tools, you can dramatically reduce the risk of a test system failure.

Parallelism as a Competitive Advantage

Traditional reliability systems have served their purpose for decades; however, the inability of these systems to provide and analyze massive amounts of reliability data is becoming a bottleneck. To address these needs, many companies are turning to modular platforms, such as PXI, to build highly parallel WLR systems with high uptime and the latest commercial processors.

Using the software-defined architecture of these systems, companies can maintain control of their intellectual property and scale their systems as requirements change. This approach satisfies their need for more reliability data at a lower cost and positions them well to address the ever-changing test requirements of the future.

Source URL: http://electronicdesign.com/test-measurement/highly-parallel-wafer-level-reliability-systems-modular-smus