KEITHLEY

WHITE PAPER

SMU-Per-Pin System Architecture Supports Fast, Cost-Effective Variation Characterization

by
Michael Wen-Ping Chao, Wayne Liao, Jerry Chiang, and Edward Kuo
Keithley Instruments, Inc., Taiwan
Hsinchu, Taiwan
email: chao_michael@keithley.com

As the knowledge of high density test structure design has grown quickly in recent years, the ability to implement fast variation characterization techniques is often limited by the number of Source-Measure Units (SMUs) available in conventional parametric testers. Overall throughput on parametric testers can be limited by the SMU resources when parallel test techniques are being implemented. In addition to the limitations imposed by the number of SMUs, the overhead involved in running multiple tests in parallel can also prove to be a bottleneck. This white paper compares different strategies for minimizing these bottlenecks and outlines a new approach that employs an SMU-based test system that provides the high throughput necessary for variation characterization while keeping the cost to semiconductor manufacturing facilities affordable.

Introduction

As device dimensions continue to scale downward, semiconductor manufacturing requires more parametric tests to provide greater insight into variations so engineers can make adequate corrections to improve overall yield. Performing more parametric tests often demands more test structures

Keithley Instruments, Inc.
28775 Aurora Road
Cleveland, Ohio 44139
(440) 248-0400
Fax: (440) 248-6168
www.keithley.com

(requiring more silicon space) and higher tester throughput. Although strategies for increasing device density (e.g., addressable array structures) without dramatically increasing the amount of silicon space needed are commonly discussed, this white paper focuses on tester throughput improvement.

Parallel test is one often-used strategy for increasing test throughput. In theory, if one were able to test four times as many devices within one probe touchdown, throughput could be increased by 4×. However, conventional SMU/switch-matrix-style parametric test systems (***Figure 1***) are often limited by the number of SMUs installed and the size of switch matrix. Doubling or tripling the size of the switch matrix to accommodate more SMUs is often infeasible because the extra leakage and parasitic capacitance that would come with a larger matrix would slow tester performance. As a result, alternative system architectures are being explored.

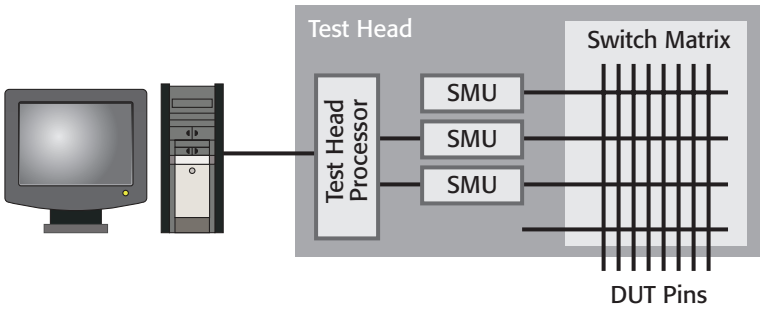


Figure 1. Although conventional parametric testers use a high speed control bus, overall throughput on high density test structures is often limited by the number of SMUs available, as well as by the number of rows (instrument pathways) in the switch matrix.

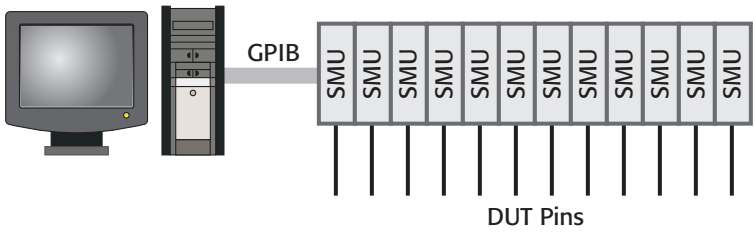


Figure 2. Alternative SMU-per-pin tester architecture

Figure 2 illustrates an SMU-per-pin system architecture that's an alternative to conventional parametric tester designs. Using this configuration, one SMU could be connected

to each of the device pads and the number of parallel test threads would be based simply on the number of devices available on the test structure.

However, for a simple architecture like this, the efficiency of the test sequences used is critical to overall tester performance. Simply linking pieces of test code together is unlikely to produce significant throughput improvement without optimization. Figure 3a illustrates where test time is spent in each step of a single-thread test sequence. Figure 3b shows the result of putting the same commands together as an example of three-thread parallel test. Although force and measure actions can be executed in parallel, the commands to each of the instruments are actually sent out sequentially due to the nature of the GPIB control bus. The communication overhead may not be negligible in the test.

(a) Single thread

(b) Multiple threads

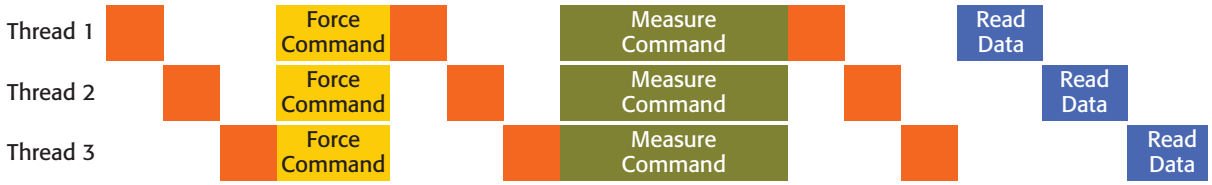


Figure 3. Communication overhead associated with control activities. (a) Time chart for a single-point force-measure test. (b) Time chart for three single-point force-measure tests executed in parallel.

Optimize Control Sequence

Even though the control bus supported on a benchtop SMU instrument usually isn't optimized for massive parallel test in terms of communication speed and synchronization (or at least not as optimized as that of a parametric tester), it's still possible to improve test throughput by leveraging other features the instruments provide. Using the test structure shown in ***Figure 4*** as an example, several strategies are available to minimize overall test time.

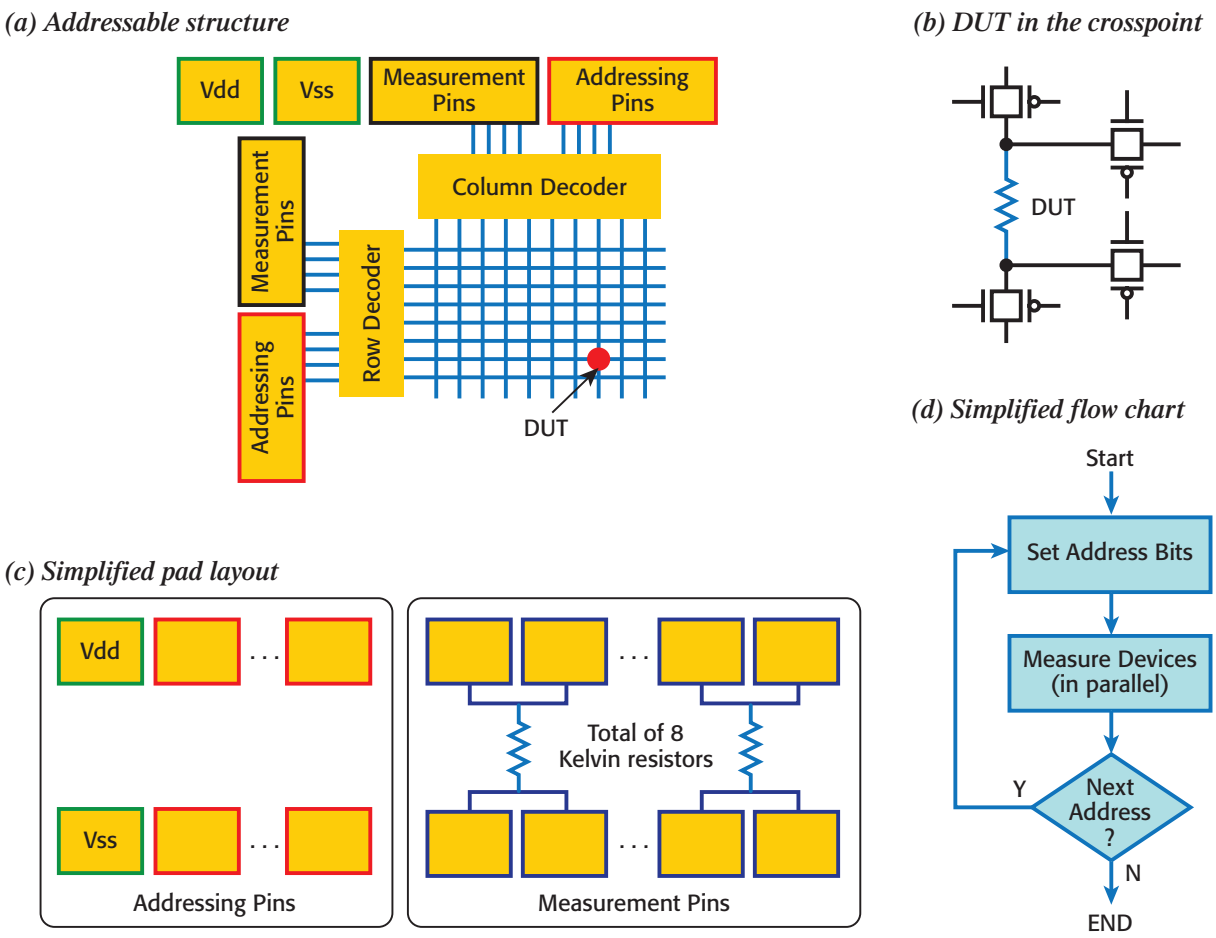
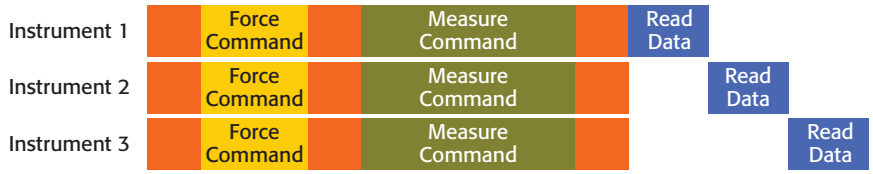


Figure 4. (a) Overview of an addressable structure. Each of the crosspoints of the row/column decoder represents one DUT (or multiple DUTs). (b) Each of the DUTs can be accessed via the transfer gate, which is controlled by the decoders. (c) Simplified pad layout of an addressable structure. Each of the address selections provides access to eight Kelvin resistors. (d) Flow chart that illustrates the procedure to test the addressable structure.

Strategy 1: Control Bus Broadcast

A GPIB bus supports broadcast to all addresses at the same time, a feature one can take advantage of when all the tests in each of the threads are the same. ***Figure 5a*** illustrates how the communication overhead from the second thread to the last one can be eliminated.

(a) Bus broadcast



(b) Distributed control code

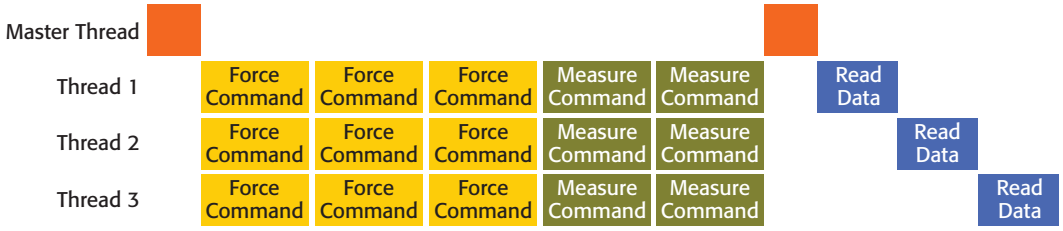


Figure 5. Methods to minimize communication overhead effectively: (a) Communication overhead (the orange blocks) may be reduced by using the broadcast method supported by GPIB. (b) Communication overhead can be further reduced when the test code can be stored in the instrument locally and the main test program sends only one command to trigger a series of force-measure activities, such as calling a function.

Strategy 2: Distributed Control Code

When a test requires some combination of force-measure actions, such as a Kelvin resistor measurement, for example (force current on force terminal and measure voltage on two sense terminals), the amount of communication overhead for each of the commands can still be significant. Some of the instruments allow storing test sequences locally in the instrument or mainframe. The system controller can simply trigger the test sequence with a single command then retrieve results after completion.

Strategy 3: Trigger via Digital I/O

The advantage of distributed control code may be further expanded through the use of digital I/O. A feedback loop can be established to trigger the next sequence if a digital output signal can be sent from the last thread (group), upon the completion of the test sequence, to the digital input of the main thread (group), which manages the test flow. Although it is difficult to predict which thread will be the last one to be completed, adding an AND gate can resolve the issue (***Figure 6***). This strategy can be used to eliminate bus communication time at the beginning of each sequence while the memory of the instrument is capable of storing all the

results in the loop. Retrieving all test results at the end of the loop also significantly reduces data read time.

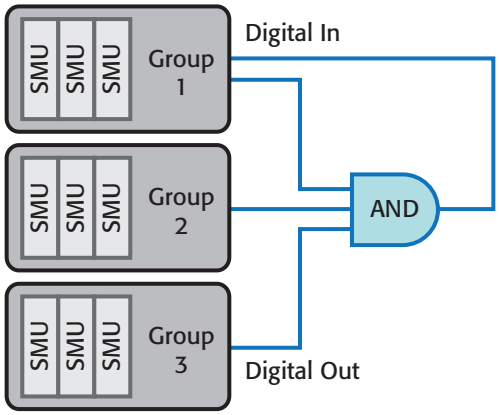


Figure 6. A hardware trigger may be implemented simply by adding a multiple input AND gate.

Strategy 4: Internal Trigger Bus

The sweep command that some newer benchtop SMUs can support can accelerate I-V curve measurements (e.g., I_DV_G sweeps); the alternative way of making the same measurement via force-measure commands in a FOR loop often takes longer due to the difference in execution efficiency between hardware and software triggers. However, an instrument or mainframe's internal trigger bus may have limitations in terms of parallel threading.

Strategy 5: Compress Result Retrieve Time

Unless the system allows all test results to be stored in the memory of the main node during the test sequence and then retrieved via a single command to the master node, the most commonly used technique is to store as many results locally (i.e., in the instrument) as possible, then allow the system controller to collect data once the whole test sequence is complete.

Addressing Control

It may be helpful to discuss how address selection can be made on the tester and how the control of address selection affects overall test time.

Conventional parametric testers don't support digital I/O at the pins, so the address selection on a parallel addressing array can only be made by connecting all high bits (pins) to an SMU while low bits (pins) are grounded. Next, address selection requires

first disconnecting all address pins before connecting high/low bits to the SMU/ground accordingly. Some experiments on an addressable Kelvin resistor structure suggest that the overhead on address selection could take twice as long as the actual force-measure action performed on the device.

If cost were no object, adding an SMU to each of the address pins would easily provide the capability of address selection. The use of a programmable power supply would be another possible alternative. However, the overhead involved in controlling a large number of SMUs or power supplies makes both methods less attractive.

Digital I/Os offer a more efficient approach too because one address selection can be made by sending an instrument a single command that programs all bits at once. This capability is readily available on many benchtop SMUs. Although most digital I/Os only support a fixed output level, which is usually too high for an addressable array structure, a simple voltage leveler can be added to recondition the digital signal to an appropriate level.

Throughput Analysis

Control Bus Architecture

A recent test time simulation suggests that both bus broadcast and distributed test code methods can achieve an 8.6× test time reduction with eight DUTs tested in parallel (not including the time necessary for data retrieval). The original method used was sequential test, which sends GPIB commands to control each of the force/measure actions one by one. Overall bus control time was reduced by 87% with both methods. The test assumes a 2ms bus control time. Note that distributed test code has an overhead dependent on the number of tests to be performed in parallel—although threads can work in parallel, they are actually created one by one. On the other hand, bus broadcast may be at a disadvantage for complex tests such as I-V sweeps. Bus broadcast also has the limitation that all units in the bus must take the same action.

Address Selection Speed

Addressing control time on a conventional parametric tester can represent as much as 30% of the overall test time because the SMU has to recreate the connection and bias on the address pins every time it moves to the next address. Even with a longer test, such as V_T test using the maximum g_m method, addressing control time may still represent 10% of the overall test time on a conventional parametric tester. In contrast, the time necessary to select an address

via digital I/O is on the order of a few milliseconds or less. With digital I/O integrated with the parametric tester, overall test time can be reduced by 15% based on the example device illustrated in ***Figure 4***.

Data Retrieval

If one focuses only on test time savings with parallel test without also optimizing the data retrieval process, the time it takes to get data back to the system controller can become significant. Although data retrieval time represents only 6% of overall test time in sequential mode, it can increase dramatically to 34% when testing eight DUTs in parallel if the data is still being retrieved point by point (sequentially). However, if it's possible to leverage the memory buffer on the SMU to read all data at the end of the test, overall throughput gain on an eight-DUT parallel test case may be improved from 6× to nearly 8×.[1]

Test Time Comparison

Recently, a simulation was conducted to assess the potential test time improvement for three different types of tests: a two-terminal resistor representing a short test (time), a Kelvin resistor representing a medium test time, and a transistor V_T (g_m max method) representing a long test. Five different system configurations were analyzed. The results are summarized in ***Table 1*** and ***Figure 7***. Note the trend for the longer test (V_T-g_m in this case) to have better parallel test efficiency when compared to the shorter test on the two-terminal resistor; it's obvious that the overhead involved in retrieving data and launching parallel test has less impact on long tests (V_T-g_m).

1. The base speed is established based on a GPIB benchtop system with a switch matrix that tests the eight-DUT Kelvin resistor array one by one and performs address selection via digital I/O; data is retrieved point by point via GPIB commands.

Table 1. Simulation results of the test time comparison.

	GPIB Benchtop System (SMU-Matrix), Addressing by Digital I/O[2]	Conventional Parametric Tester (4SMU), Addressing by SMU[3]	Conventional Parametric Tester (8SMU), Addressing by SMU[4]	Conventional Parametric Tester (8SMU), Addressing by Digital I/O[5]	Benchtop System (SMU-per-pin), Distributed Control Code, Addressing by Digital I/O[6]
8× Two-Terminal Resistors	1×	1.0×	3.7×	7.1×	7.5×
8× Kelvin Resistors	1×	1.1×	2.1×	2.4×	7.9×
8× Transistors (V_T-g_m)	1×	2.0×	5.3×	8.5×	13.7×

2. The system measures one DUT at a time, connects to the next device via switch matrix. All actions performed using GPIB commands. Test results are retrieved one by one.
3. All tests are performed sequentially, with address selection performed by the SMU (connect all high pins to SMU and low pins to GND).
4. Eight SMUs is a good number for testing four resistors, two Kelvin resistors, or three transistors in parallel (one SMU is reserved for addressing); address selection is performed by the SMU (connect all high pins to SMU and low pins to GND).
5. Eight SMUs is a good number for testing eight resistors, two Kelvin resistors, or four transistors in parallel; data from eight DUTs is retrieved together at each of the address, address selection is made with digital I/O.
6. Assume no SMU resource restriction on measurement pins; system controller triggers the test function, which is distributed and stored in the instruments; data from eight DUTs is retrieved together at each of the addresses: address selection is made with digital I/O.

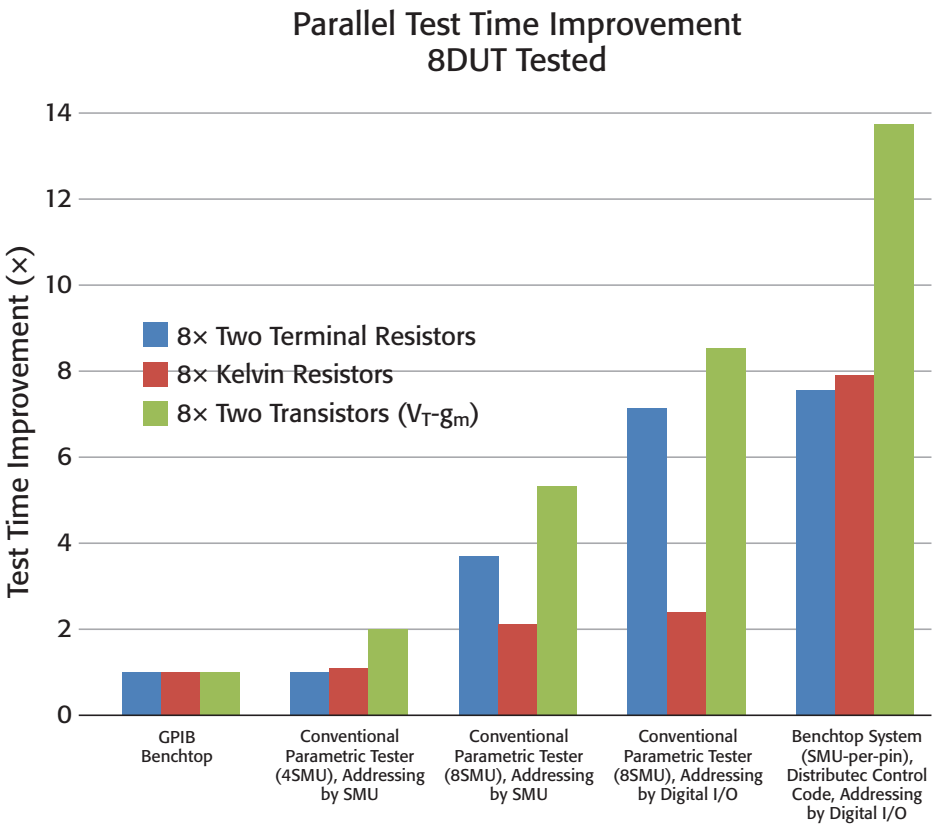


Figure 7. A graphical representation of the simulation result of the test time comparison.

System Integration

The systems illustrated in ***Figures 8*** and ***9***, both based on commercially available SMUs, are two examples of the new system architecture. Both are capable of supporting distributed control code. However, given that a parallel Kelvin resistor structure is the primary device to be measured in the project, communication overhead and data retrieval time are critical in the selection of system architecture. Due to the restriction of grouping SMUs across mainframes and the fact that test results can only be collected from the mainframes one by one, the system architecture illustrated in ***Figure 9*** is superior to the one in ***Figure 8***.

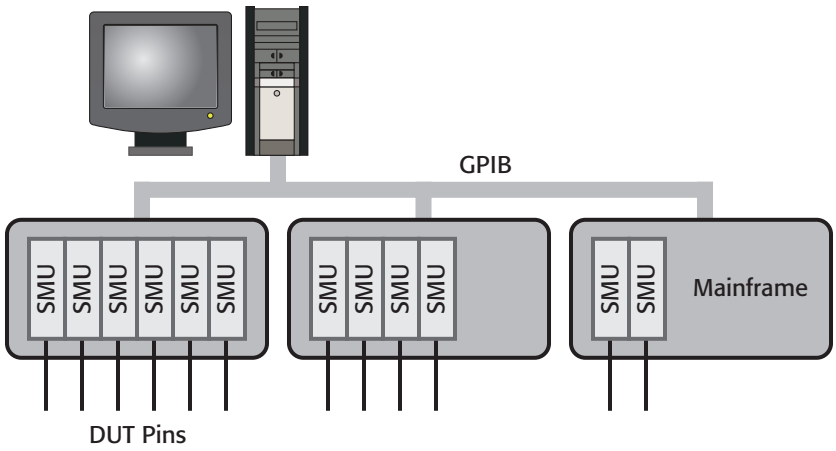


Figure 8. Alternative system integrated from benchtop mainframes. Although high speed control/ synchronization is supported within each of the mainframes, the system controller typically connects the mainframes via GPIB. Some of the mainframes available on the market are capable of storing test code locally as a function and executing the function when a trigger signal is received from a remote controller.

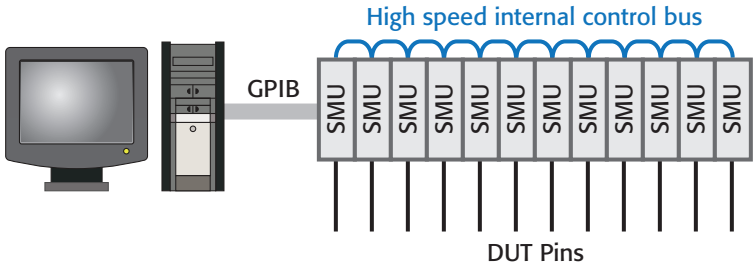


Figure 9. Alternative system integrated from benchtop SMU instruments. Some of the SMUs currently on the market support distributed control code.

The system illustrated in ***Figure 9***, which is configured using Keithley's Series 2600A System SourceMeter® instruments, integrates a total of 36 SMU channels for measurements on an addressable eight-DUT Kelvin resistor structure. Each of the SMU channels has FORCE and SENSE terminals connected all the way to the probe card to eliminate cabling-induced errors. Address selection is achieved via digital I/O from one of the SMU chassis. The system controller collects results from the master node at the end of each address. The overall cost of the system is less than that of a 36-pin conventional parametric tester. (For more information on this system development approach, see the section titled "Series 2600A System SourceMeter® Instruments and TSP® Express."

Initial results of tests performed on an addressable eight-DUT Kelvin resistor structure at an IC manufacturer site indicate that a 7× throughput advantage over the speed of a conventional parametric tester is possible. Using an SMU-per-pin system configuration, test sequences that once required 40 hours to complete were completed in less than six hours.

Conclusion

Methods for achieving optimal throughput have been discussed, and a system has been built to prove that SMU-per-pin system can dramatically increase test throughput on an addressable eight-DUT Kelvin resistor structure without significantly increasing the cost.

The throughput gains made possible through the use of SMU-per-pin systems would make it possible to complete many variation characterization tasks within one day. Such an increase in efficiency could soon accelerate the adoption of variation characterization as part of the in-line test process.

References

[1] Christopher Hess *et al.*, "High Density Test Structure Array for Accurate Detection and Localization of Soft Fails," *IEEE Conference on Microelectronic Test Structures*, March 2008.

[2] Brad Smith *et al.*, "A Novel Biasing Technique for Addressable Parametric Arrays," *IEEE Conference on Microelectronic Test Structures*, March 2008.

[3] Muthu Karthikeyan *et al.*, "Short-Flow Test Chip Utilizing Fast Testing for Defect Density Monitoring in 45nm," *IEEE Conference on Microelectronic Test Structures*, March 2008.

[4] Karen M. G. V. Gettings and Duane S. Boning, "Test Circuit for Study of CMOS Process Variation by Measurement of Analog Characteristics," *IEEE Conference on Microelectronic Test Structures*, March 2007.

[5] Xiaoju Wu *et al.*, "Impact of Sinter Process and Metal Coverage on Transistor Mismatching and Parameter Variations in Analog CMOS Technology," *IEEE Conference on Microelectronic Test Structures*, March 2007.

[6] Kelvin Yih-Yuh Doong *et al.*, "Field-Configurable Test Structure Array (FC-TSA): Enabling Design for Monitor, Model, and Manufacturability," *IEEE Transactions on Semiconductor Manufacturing*, May 2008, Vol. 21, No. 2.

[7] Naoki Izumi *et al*., “Evaluation of Transistor Property Variations Within Chips on 300-mm Wafers Using a New MOSFET Array Test Structure,” *IEEE Transactions on Semiconductor Manufacturing*, August 2004, Vol. 17, No. 3.

[8] Robert Lefferts Ph.D., Chris Jakubiec, “An Integrated Test Chip for the Complete Characterization and Monitoring of a 0.25µm CMOS Technology that Fits into Five Scribe Line Structures 150µm by 5,000µm,” IEEE Int. Conference on Microelectronic Test Structures, March 2003.

[9] Tetsuo Chagawa, “Measurement of the MOSFET Drain Current Variation Under High Gate Voltage,” *IEEE Conference on Microelectronic Test Structures*, March 2008.

Series 2600A System SourceMeter® Instruments and TSP® Express

Keithley's Series 2600A System SourceMeter instruments are high performance I-V source-measure instruments designed for use either as bench-top I-V characterization tools or as building block components of multi-channel I-V test systems. Each Series 2600A SourceMeter instrument combines a precision power supply, a true current source, a DMM, an arbitrary waveform generator with measurement, an electronic load, and a trigger controller – all in one instrument. They offer an ideal solution for I-V functional test and characterization of a wide variety of semiconductors, materials, and electronic devices. The Series 2600A family includes six different models, in single- or dual-channel versions, and a wide dynamic range of 1fA to 10A and 1μV to 200V. The dual-channel (1fA, 10A pulse) Model 2636A is widely used in semiconductor test applications that require exceptional current measurement sensitivity.

Each Series 2600A instrument has an embedded Test Script Processor (TSP®) that allows the instrument to run complete test programs (scripts) right on the instrument. Because the test scripts can contain any sequence of routines that can be executed by conventional programming languages (including decision-making algorithms), this feature allows entire tests to be managed by the instrument without the need to send readings back to the PC for decision making. This means that delays due to GPIB traffic congestion are eliminated and overall test times are greatly improved. TSP technology also supports "mainframe-less channel expansion." The TSP-Link channel expansion bus (which uses a 100 Base T Ethernet cable) allows multiple Series 2600A instruments (and other TSP instruments) to be connected in a master-slave configuration and operate as one integrated system. TSP-Link supports up to 32 units or 64 SMU channels per GPIB or IP address. This allows for an all-but-unlimited channel count, allowing users to scale their systems to fit their specific applications. All Series 2600A instruments include a built-in TSP Express software tool that allows users to perform common I-V tests quickly and easily without programming or installing software.

For applications that require tight instrument synchronization, a high performance, hardware-driven trigger model allows timing at each I-V source-measure step to be controlled precisely and operation between SMU channels and/or external instrumentation to be synchronized at hardware speeds less than 500 nanoseconds. This represents a 400× improvement in precision timing compared to previous solutions, allowing more tightly controlled test conditions than ever before.

TSP Express provides an intuitive user interface to set up and run basic and advanced tests easily, including nested step/sweeps, pulse sweeps, and custom sweeps for device characterization applications. For applications where single point I-V source-delay-measure is all that is needed, TSP Express also provides an interface to configure and measure discrete points quickly. TSP Express data can be viewed in graphical or tabular formats and can be readily exported to a .csv file for use with spreadsheet applications. An automatic script generation feature simplifies the process of writing custom programs for more advanced test requirements.

The Series 2600A's parallel testing capability allows all SMU channels to run the same or different tests on any number of channels in the system synchronously or asynchronously. Additionally, the Series 2600A system can be dynamically reconfigured via software – without rewiring the test system – for maximum flexibility and efficiency when testing a wide mix of devices.

All Series 2600A instruments are LXI Class C compliant and feature a built-in LXI Web interface, which makes it simple to configure measurements and transfer data at high speed. Remote testing and monitoring for troubleshooting are quick and easy. All the capabilities needed to start testing with Series 2600A instruments are instantly accessible via the Web interface. A Quick Start menu provides an intuitive point-and-click environment for common instrument functions; a Tools menu offers project utilities and advanced features for writing custom scripts for creating more sophisticated test sequences.

KEITHLEY

A GREATER MEASURE OF CONFIDENCE

KEITHLEY INSTRUMENTS, INC. ■ 28775 AURORA ROAD ■ CLEVELAND, OHIO 44139-1891 ■ 440-248-0400 ■ Fax: 440-248-6168 ■ 1-888-KEITHLEY ■ www.keithley.com

BELGIUM
Sint-Pieters-Leeuw
Ph: 02-3630040
Fax: 02-3630064
info@keithley.nl
www.keithley.nl

CHINA
Beijing
Ph: 8610-82255010
Fax: 8610-82255018
china@keithley.com
www.keithley.com.cn

FINLAND
Espoo
Ph: 358-40-7600-880
Fax: 44-118-929-7509
finland@keithley.com
www.keithley.com

FRANCE
Saint-Aubin
Ph: 01-64532020
Fax: 01-60117726
info@keithley.fr
www.keithley.fr

GERMANY
Germering
Ph: 089-84930740
Fax: 089-84930734
info@keithley.de
www.keithley.de

INDIA
Bangalore
Ph: 080-26771071, -72, -73
Fax: 080-26771076
support_india@keithley.com
www.keithley.com

ITALY
Peschiera Borromeo (Mi)
Ph: 02-5538421
Fax: 02-55384228
info@keithley.it
www.keithley.it

JAPAN
Tokyo
Ph: 81-3-5733-7555
Fax: 81-3-5733-7556
info.jp@keithley.com
www.keithley.jp

KOREA
Seoul
Ph: 82-2-574-7778
Fax: 82-2-574-7838
keithley@keithley.co.kr
www.keithley.co.kr

MALAYSIA
Penang
Ph: 60-4-643-9679
Fax: 60-4-643-3794
chan_patrick@keithley.com
www.keithley.com

NETHERLANDS
Gorinchem
Ph: 0183-635333
Fax: 0183-630821
info@keithley.nl
www.keithley.nl

SINGAPORE
Singapore
Ph: 65-6747-9077
Fax: 65-6747-2991
koh_william@keithley.com
www.keithley.com.sg

SWEDEN
Stenungsund
Ph: 08-50904600
Fax: 08-6552610
sweden@keithley.com
www.keithley.com

SWITZERLAND
Zürich
Ph: 044-8219444
Fax: 044-8203081
info@keithley.ch
www.keithley.ch

TAIWAN
Hsinchu
Ph: 886-3-572-9077
Fax: 886-3-572-9031
info_tw@keithley.com
www.keithley.com.tw

UNITED KINGDOM
Theale
Ph: 0118-9297500
Fax: 0118-9297519
info@keithley.co.uk
www.keithley.co.uk

Printed in the U.S.A. No. 3032 03.24.09