Correlating Models and Silicon for Improved Parametric Yield

Rob Aitken[1], Greg Yeric[2], David Flynn[3]
1=ARM Inc., San Jose CA
2=ARM Inc., Austin TX
3=ARM Ltd., Cambridge UK
rob.aitken@arm.com

***Abstract*— This paper discusses one of the key challenges of design-for-yield: namely, the difficulty in correlating observed behavior with modeled behavior. In order to achieve good parametric yield, the design process must account for a large number of sources of variability in the silicon, ranging from those inherent in the device and wire models themselves through approximations made in library modeling, extraction, tool algorithms and so on. The problem is further complicated by defects and systematic errors that can be present in early silicon but are expected to be fixed as part of the volume ramp. In addition, environmental factors such as temperature and power delivery must be understood, and variation in the measurement equipment must also be correctly accounted for. Examples are given for validating standard cell and memory based designs as well as a general methodology that can be used to enable chip bring-up.**

Keywords – yield optimization, variability, silicon correlation

I. INTRODUCTION

Nanometer-era process complexities can create context-dependent behaviors which result in significant differences between the performance of process monitor test structures and the actual devices used in an SoC (System on Chip) design. Stress effects, implant proximity effects, lithographic non-idealities, etch and chemical-mechanical polish (CMP) density effects, and many others, can create model/testchip/hardware differences which can lead to a loss of yield understanding and/or sub-optimal design [1][2]. Even with the inclusion of restrictive design rule sets, a foundry cannot practically monitor the large number of electrically unique device contexts which could be used by its various customers, so it becomes incumbent on the design organizations to monitor specific key contexts, their differences from standard models and from process monitor test structure results, and the effects of these differences on key standard cell and/or SoC design metrics.

At the same time, it is often not feasible for a fabless design organization to generate correct and timely silicon characterization data via dedicated process monitor test chips. This problem is particularly acute for fabless design groups operating in early or even pre-release stages of technology nodes, where process definition and centering is still ongoing and changes in process characteristics can be anticipated between a process characterization test chip snapshot and SoC production runs.

To meet these needs, we have created a test qualification vehicle architecture which includes process, voltage, and temperature monitoring elements, library validation, memory structures, and a microprocessor based control system. In this paper we discuss process monitors in the form of ring oscillators, to be seamlessly inserted into a synthesizable microprocessor design. Due to their inherent monitoring of process performance and their digital I/O nature, ring oscillators (ROs) have been heavily utilized for dedicated test chips designed for thorough process monitoring [3][4]. Recently, small sets of ROs have been directly embedded into CPUs so that local variability can be more accurately characterized within a CPU context [5]-[7]. In addition, it is desirable to predict the performance of complex CPUs early in a process generation; we present a set of structures that are able to predict processor performance and demonstrate their effectiveness on a commercial 32nm process.

Parameter (Scale factor = α)	Dennard Scaling	Scaling Now
Dimensions	$1/\alpha$	$\sim 1/\alpha$
Oxide thickness	$1/\alpha$	1
Voltage	$1/\alpha$	1
Drive Current	$1/\alpha$	$1/\alpha$
Capacitance	$1/\alpha$	$1/\alpha < C < 1$
Power/Circuit	$1/\alpha^2$	$1/\alpha$
Power Density	1	α
Delay/Circuit	$1/\alpha$	~1

Table 1. Scaling challenges

978-3-9810801-7-9/DATE11/

II. VARIABILITY AND MARGINS

The basics of CMOS scaling, as outlined in the classic paper by Dennard et al of IBM [8], are shown in Table 1. This scaling served the semiconductor industry well for nearly 30 years, but physical limitations have slowed conventional scaling. First of all, voltage scaling slowed and has nearly stopped at around 1V. As can be seen in the table, the lack of voltage scaling means that delay and dynamic power scaling stops. This also lowers power dissipation per circuit, which results in power density increasing, rather than staying constant. Oxide thickness was next to slow down – as oxides shrunk to single digits of atoms thick, additional shrinking was simply not possible. However, economic pressure to continue scaling did not abate, and so alternative approaches were required.

Some of these are new materials. For example, moving to high-k metal gate (HKMG) transistors provides a one-time "bump" in benefits previously obtained by scaling gate oxide even when the oxide doesn't scale. Other such techniques include providing for multiple dopant concentrations and strain engineering in the silicon.

Additionally, even though device dimensions continue to scale, the mechanics of doing so has become increasingly complex. The wavelength of light used for photolithography has stopped scaling at 193nm even as feature sizes continue to shrink. The added complexity of the photolithography and mask making process has led to more complex design rules, meaning that the simple pattern shrinking envisaged by Dennard has been impossible since the 90nm silicon generation (approximately 2002).

Even with all of the technical fixes being used in the process of fabricating circuits, consumer demands for scaling can no longer be achieved without design changes. Significant standard cell and memory architecture changes are made at each process node in order to squeeze additional performance, area, or power out of devices that are increasingly reluctant to provide the gains. Further, parasitic resistance and capacitance are rapidly becoming dominant effects that must be considered when evaluating performance. The additional complexity is reflected in the increasing cost of developing each node.

Finally, shrinking dimensions, increasing numbers of devices, and processes that increasingly bump up against physical walls has led to increasing concern about variability. Variability is now omnipresent: individual transistors vary in their dimensions and performance, within chips and among chips, by 10% in dimension and 50% in performance. Wire capacitances and resistances can easily vary by 30% or more between chips and even between adjacent metal layers. On-chip variation in voltage and temperature can further exacerbate these problems. The standard solution to this variation is to add margin to a design, which further limits its ability to scale. High performance design teams need to quantify and account for margins in order to get the gains that design teams 15 years ago could achieve simply by switching to a new process. In order to accomplish this, leading edge design teams must spend more effort prior to production design in measurement and analysis of intrinsic variability. In section III we describe one of our efforts.

III. MEASUREMENT STRUCTURES

The overall measurement system is a microprocessor based system intended for two primary roles: both a minimal ARM CPU technology demonstrator and also an on-chip embedded test controller. The platform is named "MicroSTEP", where the STEP naming stands for

- System – reference technology demonstrator
- Test controller – support software-based "BIST" and data logging
- Evaluation vehicle - exhibition quality demonstrator with OLED display panel drive
- Platform – extensible memory-mapped MCU, able to host multiple-CPUs

The approach is shown in Figure 1. It is a clean simple design with minimal constraints and STA complexities. The small, simple microcontroller design can be built with basic Standard Cell library and basic digital I/O pads, so can built on early technology long before PLL, DDR pads or even RAM compilers exist. The RTL is designed for portability, allowing for FPGA prototyping in the first instance. Several designs have been taped out in sub-32nm technologies.

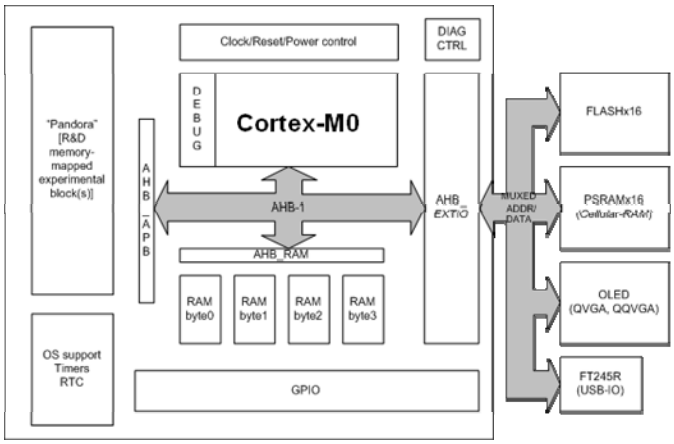

Figure 1. MicroSTEP block diagram

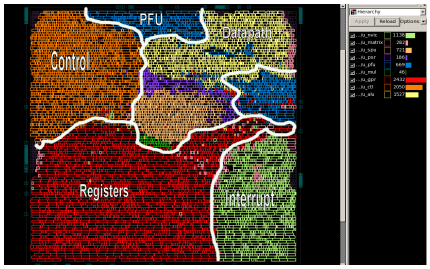

Figure 2. Cortex-M0 layout at 32nm (120um x 120um)

The configurable aspects of the design allow for a variety of extensions, ranging from ring oscillator blocks [2] to full CPUs. Each of these can be accessed via AMBA AHB interconnect in addition to any dedicated pins [12].

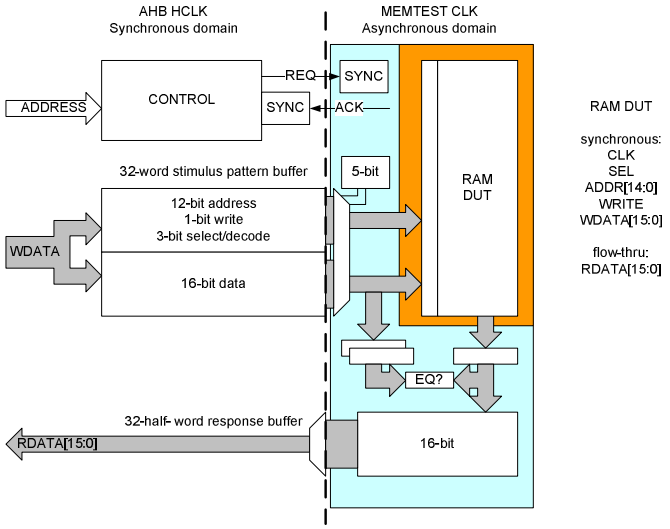

Figure 3. Multiple test access methods

Figure 3 shows an example of using the MicroSTEP architecture to drive a memory self test. The memory clock is allowed to be asynchronous to the AHB clock, again for flexibility. Functional tests can be applied via a 32 bit access port using a standard AHB protocol. In addition, an 8 bit protocol is defined allowing for a USB-based access when the chip is mounted on a demonstrator board. This provides for a very low cost and highly configurable method for comprehensive data gathering using PC-based applications, which facilitates custom test generation and application, especially for phenomena that are new or were unanticipated when the test chip was first designed. In addition, the architecture supports a standard memory BIST access, including pipelined access features that are useful for efficient embedded memory test in microprocessors.

Overall, the architecture is designed for flexibility, enabling several standardized physical access mechanisms to test structures whether via ATE at wafer or package or dedicated hardware, plus software access via the central microcontroller, while not precluding custom dedicated access for individual structures.

IV. Predictive Behavior

A key requirement of the system is the ability to predict power, performance, and area (PPA) numbers for future products in a given technology. This requires both correct design of predictive test structures as well as a method of correlating silicon results to PPA numbers and enabling effective design margins.

An ideal predictive structure for CPU performance would precisely measure critical paths on an implemented core. This is challenging for several reasons. First is size: a full application processor is a large complex design and this design process is fundamentally incompatible with an early life cycle silicon process. Second, processes evolve significantly from their early implementations: devices change, parasitics change, yields improve, variations decrease, and so on. Third, early silicon comes with models with limited silicon backing (e.g. devices, interconnect), making the design infrastructure needed for a high performance design (e.g. I/O, PLL) very challenging. Finally, the environment on a test chip is very different from "real" silicon, including temperature control, packaging, and power delivery.

Because of these difficulties, it is infeasible to use the actual product. Simple device measurements (Ion/off) will give some predictive data. Increasing complexity can lead to better results: Ring oscillators will give more information, sub-blocks (e.g. ALU) will give more, simple processors will give more still. However, this information can be misleading if the structures are not designed correctly.

We have found the following structures, among others, to be effective performance and area predictors:

For standard cell libraries

- Ring oscillators, especially those varying one or two parameters from a reference design and those measuring the tradeoffs between device DC behavior and actual AC performance [2]
- Multiple track height cells
- Placed and routed blocks

For embedded SRAM

- Sample instances
- Sense amplifier measurement structures
- Bit cell measurement structures
- Timing measurement structures
- Variability measurement structures, especially those that enable "tail bit" information [9]

For processors

- CPU path oscillators [2]
- Sample blocks
- Multiple microcontroller implementations using different libraries
- Simplified (reduced instruction set) processor implementations [10]

Accurate power measurement requires dedicated power supply access to the relevant structures, and typically also involves substantial over-design of the power network due to uncertainty and implementation challenges in metal.

Taken individually, none of these is able to predict PPA accurately, but by judiciously combining data, a reasonably complete picture can be obtained.

V. Experimental Results

The method described in this paper has been implemented on a number of test chips at and below 40nm. A complete cycle has been observed at 32nm and some of the results are reported here.

Our process technology qualification vehicle architecture at the 32nm node includes many of the structures described in section IV. Many results are detailed in [2]. The oscillator block contained 256 oscillators of 90 different varieties, including some dedicated to key library parameters such as gate pitch and device channel length, as well as to quantities important for library characterization and margining, such as well proximity effect. Figure 4 shows several key learnings, including limited performance and variability cost for slightly increased device length, showing an excellent tradeoff for leakage and performance, and also clearly shows that the well edge should be skewed towards P devices (wpe_p) rather than N devices (wpe_n), as expected.

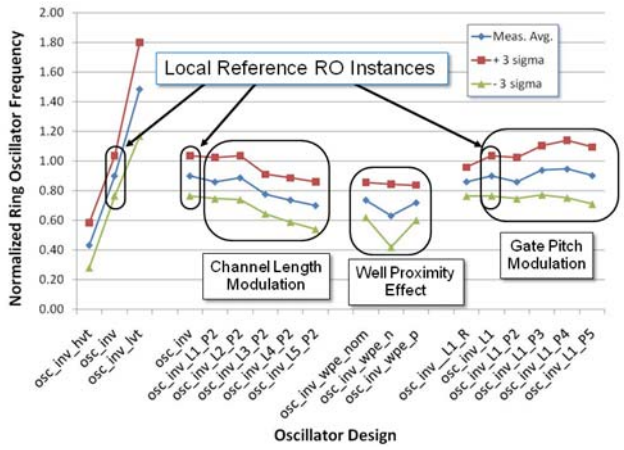

Figure 4. Oscillator measurements for various library structures

The design also included some dedicated CPU path oscillators. These were designed to mimic key performance paths in a high performance applications processor, and were composed of a representative set of standard cells rather than simple inverter based structures shown in Figure 4. Initial measurements on the first silicon available showed very promising results, suggesting that GHz range processors were feasible in the particular 32nm technology being observed. These results are tabulated in Figure 5.

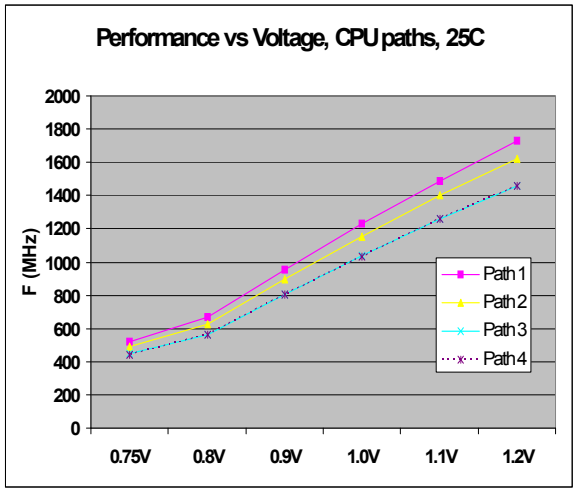

Figure 5. Early CPU path results

These results have been corroborated by subsequent implementation of a high performance CPU at the 32nm node [11]. Figure 6 shows the relationship between measured predictive silicon results (averaging all 4 paths), design sign-off values, and measured performance of the final implemented CPU core. It can be seen that the predictive data does correlate strongly with final implementation, demonstrating the usefulness of the approach.

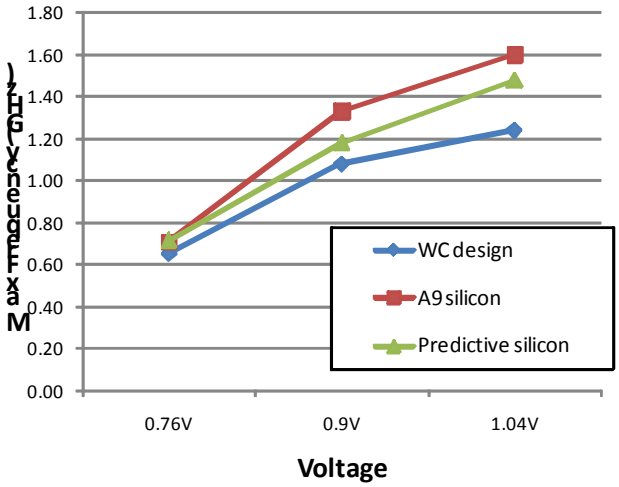

Figure 6. Relative performance of worst case design (WC), predictive silicon, and final core (A9)

VI. Conclusions

We have presented a test qualification vehicle designed to enable effective correlation of predictive silicon measurements with final product performance. The architecture continues to evolve, but currently is based on the MicroSTEP platform, incorporating a microcontroller together with a standard bus, enabling a variety of access methods for test application and data gathering, as well as acting, in and of itself, as a technology demonstrator and PPA predictor.

ACKNOWLEDGMENTS

The authors would like to thank John Biggs, David Bull, Vikas Chandra, Brian Cline, Shidhartha Das, Betina Hold, David Howard, Sachin Idgunji, Imran Iqbal, James Myers, David Ondricek, David Pietromonaco, Cezary Pietrzyk, Bal Sandhu, and others for their significant contributions to this project.

REFERENCES

[1] C. Hou, "Design and Process Co-optimization for 28nm/22nm and Beyond – A Foundry's Perspective" IEEE International Devices Meeting, 2009, page 449.

[2] S. Idgunji, V. Chandra, C. Pietrzyk, I. Iqbal, R. Aitken, G. Yeric, "An embedded process monitor test chip architecture", IEEE International Conference on Microectronic Test Structures, 2010, pp. 122-127.

[3] S. Ohkawa, M. Aoki, H. Masuda, "Analysis and Characterization of Device Variations in an LSI Chip Using an Integrated Device Matrix Array", IEEE International Conference on Microectronic Test Structures, 2003, pages 70-75.

[4] Rigaud, F.; Portal, J.M.; Dreux, P.; Vast, J.; Aziza, H.; Bas, G.; "Fast Embedded Characterization of FEOL Variations in MOS Devices", IEEE International Conference on Microelectronic Test Structures, 2009, pages 205-208.

[5] A. Gattker, M. Bhushan and M. B. Ketchen; "Data Analysis Techniques for CMOS Technology Characterization and Product Impact Assessment:, IEEE Interntional Test Conference, 2006, Lecture 3.3, pages 1-10.

[6] C. Cho, D. Kim, and J. Kim, "Cell Broadband Engine Performance Benchmark in 65nm SOI CMOS with Spatial, Temporal, and Parametric Process Variability Model", IEEE Asian Solid-State Circuits Conference, 2008, pages 21-24.

[7] S. Mitra, E. Volkerink, E. J. McCluskey, and S. Eichenberger, "Delay defect screening using process monitor structures", IEEE VLSI Test Symposium, April 2004, pp. 43-48.

[8] R. Dennard et al, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions", IEEE Journal of Solid-State Circuits, vol. SC-9, no. 5, pp. 256–268, October 1974.

[9] R. Aitken, A. Singhee, R. Rutenbar, "Extreme Value Theory: Application to Memory Statistics", in Extreme Statistics in Nanoscale Memory Design, Springer 2010.

[10] D. Bull et al, "A Power-Efficient 32b ARM ISA Processor Using Timing-Error Detection and Correction for Transient-Error Tolerance and Adaptation to PVT Variation", Proc. Int. Solid State Circuits Conf., pp. 284-286, Feb. 2010.

[11] "ARM Announces 32nm Cortex-A9 Processor Optimizations", ARM press release, Nov. 9, 2010.

[12] D. Flynn, "AMBA: Enabling Reusable On-Chip Designs", IEEE Micro, Vol. 17, No. 4, July 1997