Random Sampling for On-Chip Characterization of Standard-Cell Propagation Delay

S. Maggioni A. Veggetti A. Bogliolo[STI] L. Croce
STMicroelectronics
Agrate (MI) – Italy
ste34@libero.it, andrea.veggetti@st.com, luigi.croce@st.com

STI – University of Urbino
Urbino (PU) – Italy
alessandro.bogliolo@uniurb.it

Abstract

We present a methodology for on-chip characterization of the pin-to-pin propagation delay of single standard cells. A periodic waveform is provided to an input pin of the standard cell under characterization, while keeping all other inputs at non-controlling logic values. Simultaneous random sampling is then applied to input and output periodic waveforms, and propagation delay measures are obtained from the joint signal probabilities of the samples.

The proposed technique is suitable for on-chip implementation because it is simple and it doesn't require timing-accurate control signals. On the other hand, on-chip measurements can be applied to a large number of cells working in different operating conditions, providing valuable information for characterizing and validating timing models.

A test chip has been realized in a 0.18μm embedded NVM CMOS technology, and used to monitor the sub-nanosecond timing behavior of a standard cell library during process development.

1. Introduction

The timing closure is one of the most challenging issues in the design of today's integrated circuits [1,2] because of timing uncertainties due to: the lack of routing information during the early stages of a design flow, the lack of accurate timing models for logic components and interconnects and the effect of process variations on propagation delays [3]. Dealing with timing uncertainties imposes the use of conservative delay models and large time margins that lead to over-sized designs and make it difficult to reach the performance limit of the target technology [2]. On the other hand, aggressive design styles that make use of tight time margins may incur in timing violations that impose expensive trial and error iterations. As a matter of fact, the quality of a design is strongly affected by the accuracy of the timing models available for its building blocks.

In general, special test structures (TEGs) are used to measure and characterize device parameters, while timing models for library cells are obtained from electrical-level simulations annotated with layout-extracted parameters [4]. Timing models are then validated on representative test structures by comparing delay estimates with real-world measurements.

There are three main approaches to measure on-chip delays: using external probes to contact the signals of interest, using e-beams to observe electrical signals, using on-chip measurement techniques [5]. The main drawbacks of probes are intrusiveness and handling complexity, that limit the number and the accuracy of the delay measures. E-beam techniques, on the other hand, are much less instrusive, but much more expensive, so that their practical application is limited as well. Both limitations are overcome, in principle, by on-chip measurement techniques. Integrating on the same die both the device under test (DUT) and the measurement equipments has several advantages: i) measurement equipments can be placed as close as possible to the test structure, thus improving the signal to noise ratio, ii) delay measures can be performed in the actual operating conditions, iii) many test structures can be integrated on the same die, iv) a large number of measures can be simultaneously performed, v) only low-cost external instrumentation is required. On the other hand, bringing measurement equipments on chip is not a trivial task [5,6].

2. On-chip delay measures

The typical test structure used for delay measures is a delay chain, i.e., a chain composed of several cascaded copies of the cell under test. The reason for using a delay chain instead of a single cell is time resolution: the propagation delay of a single cell is often shorter than the time resolution of the measurement equipment, and it is typically shorter than local wiring delay. Cascading N cells with minimum interconnects has the effect of amplifying the delay of interest, that can then be obtained by dividing by N the measured delay. The result is the average pin-to-pin propagation delay computed over the N instances of the library cell of interest. The inherent average computation is the main limitation of delay chains, that make it impossible to measure single-cell contributions. Hence, they are suitable for typical-case characterization, but they do not provide enough information about corner cases. Moreover, if the cell of interest implements an inverting function, the signal that

Proceedings of the Fourth International Symposium on Quality Electronic Design (ISQED'03)
0-7695-1881-8/03 $17.00 © 2003 **IEEE**

propagates through a delay chain has both rising and falling transitions, making it impossible to independently characterize rising and falling propagation delays.

A delay chain can also be used as part of a ring oscillator in order to obtain delay measures from simpler frequency measures. In this case, rising and falling transitions are always measured together since they are both involved in determining the oscillation period.

In this paper we propose a new approach to on-chip delay measures that has two main advantages over previous techniques: it measures single-cell delay, it discriminates between rising and falling propagation delays.

3. The proposed approach

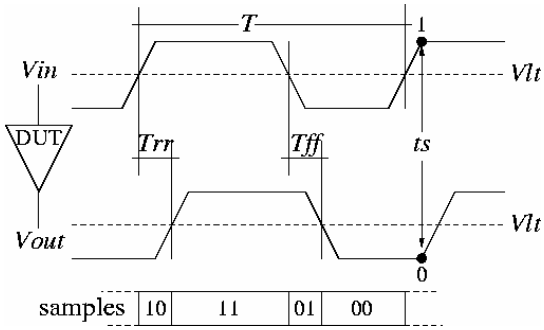

Figure 1. Illustration of the proposed approach.

Figure 1 shows a generic non-inverting standard cell (denoted by DUT, i.e., *device under test*) with a periodic input signal *Vin* that gives rise to a periodic output signal *Vout*. Input and output waveforms are represented in Figure 1 with a misalignment that represents the delay of *Vout* with respect of *Vin* (i.e., the pin-to-pin propagation delay of the cell). We sample the values of *Vin* and *Vout* at a random time instant *ts* and we digitize the sampled values according to a given logic threshold *Vlt*. What we obtain are two Boolean random variables *Sin* and *Sout*, respectively. It can be easily demonstrated that the joint probability Pr(*Sin*=1,*Sout*=0) is proportional to the propagation delay of a rising edge through the cell. In fact, the longer the propagation delay, the higher the probability of sampling the two waveforms when the input signal is already high and the output signal is still low. In particular, Pr(*Sin*=1,*Sout*=0) is equal to the ratio between the propagation delay of a rising transition *Trr* and the period of the input waveform *T*:

$$\Pr(Sin = 1, Sout = 0) = \frac{Trr}{T} \qquad (1)$$

Similarly, the joint probability Prob(*Sin*=0,*Sout*=1) is proportional to the falling propagation delay *Tff*:

$$\Pr(Sin = 0, Sout = 1) = \frac{Tff}{T} \qquad (2)$$

For inverting DUTs, propagation delays from input rising transitions to output falling transitions (*Trf*) and from input falling transitions to output rising transitions (*Tfr*) are proportional to Pr(*Sin*=1,*Sout*=1) and Pr(*Sin*=0,*Sout*=0), respectively.

The proposed approach consists of performing on-chip random sampling to measure propagation delays from signal probabilities. For the sake of simplicity, let's focus on the rising propagation delay of a non-inverting cell (*Trr*), that can be obtained from Prob(*Sin*=1,*Sout*=0) by inverting equation (1):

$$Trr = \Pr(Sin = 1, Sout = 0)T \qquad (3)$$

The key issue to be addressed is the estimation of Pr(*Sin*=1,*Sout*=0). The joint signal probability is a parameter of the parent population of a random variable that can be estimated from a random sample of *N* independent elements by computing the relative frequency of event *Sin*=1,*Sout*=0, hereafter denoted by *F*(1,0). In its turn, *F*(1,0) can be computed from event counts:

$$F(1,0) = \frac{N(1,0)}{N} \qquad (4)$$

so that the propagation delay of interest can be obtained from event counts:

$$Trr \cong \frac{N(1,0)}{N} T \qquad (5)$$

According to the theory of random sampling, the higher the value of *N*, the higher the quality of *F*(1,0) as an estimator for Pr(*Sin*=1,*Sout*=0), since its standard error is inversely proportional to the square root of *N* [7]. To this purpose it is worth noting that using periodic signals

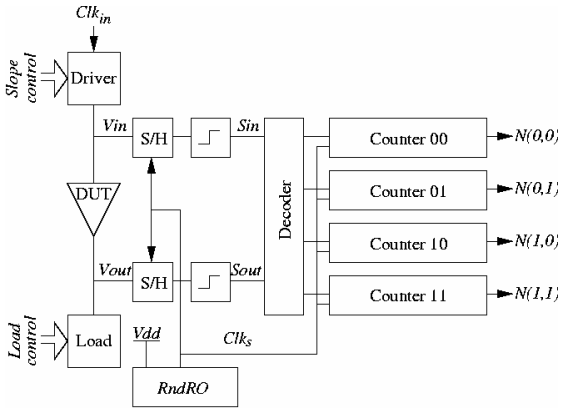

Figure 2. Schematic representation of the proposed approach.

allows us to perform random sampling over several periods, thus improving the accuracy independently of the

sampling rate. Averaging over multiple periods has the further advantage of filtering the random noise possibly due to clock jitter.
On the other hand, only independent samples give an actual contribution to the reduction of the standard error. In the worst case of sampling points with the same periodicity *T* of signals *Vin* and *Vout*, the quality of the estimator would depend only on the number of samples in a single period. In this case, extending the sampling over several periods would have the only effect of filtering signal jitter.

4. Implementation

A schematic representation of a possible implementation of the proposed technique is shown in Figure 2. Both the driver and the load of the DUT are parameterized, in order to emulate different operating conditions. In particular, the driver is composed of three parallel tri-state buffers (of size 1, 2 and 4) with independent control signals. The binary configuration of the three control signals (denoted by *Slope control* in Figure 2) selects the slope of *Vin* among the 7 available (the 000 configuration is not allowed since it wouldn't provide any signal). The input of the driver is a periodic signal (Clk_{in}) directly taken from a primary input pin.
The variable load is realized by means of two transfer gates connecting the output of the DUT to two MOS capacitors of size 1 and 2. The actual load is selected by the binary configuration of the two control signals denoted by *Load control*.
Signals *Vin* and *Vout* are sampled by two S/H circuits and compared with a common logic threshold *Vlt* to obtain logic signals *Sin* and *Sout*. The configuration of (*Sin*,*Sout*) is the input of a decoder that generates mutually-exclusive enable signals for 4 counters, clocked by the same signal

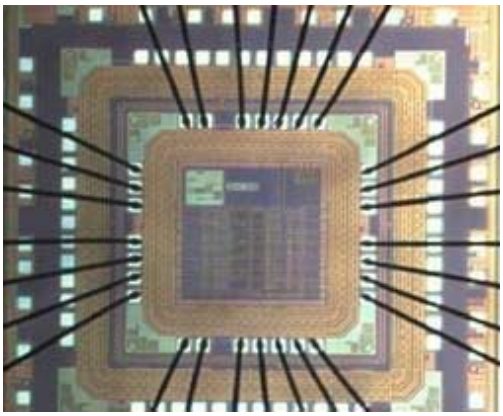

Figure 3. Top-level picture of the test chip.

used to control the S/H's.

At the end of the experiment, the four counters contain the event counts needed to compute the propagation delay across the DUT. In particular, the sum of the final values of the 4 counters is the total number of samples *N*.
A test chip implementing the scheme of Figure 2 was realized in a 0.18µm technology to measure the propagation delay of 32 different elements of a core library (including inverters, 2-input logic gates, and level-sensitive latches). The 32 DUTs were integrated on the same die, sharing I/O pins and some control circuitry. In particular, the components directly connected to the DUT (i.e., driver, receiver, S/H) were replicated, while internal random clock generators, counters and I/O buffers were multiplexed and shared. The 3 least-significant bits of the counters were also replicated for performance reasons. Each DUT was assigned with a binary address to be specified at setup time to select the target device under measure.
The top-level picture of the test chip implementing the array of 32 DUTs and their measurement equipments is shown in Figure 3. Shared circuitry (for signal generation, control and counters) takes a silicon area of 180x180µm^2, while the circuitry dedicated to each DUT requires 120x18µm^2. We remark that only a few pins are used to connect to the external test equipments, thanks to the scan-path methodology used for providing measure control configurations and for accessing results.

4.1. Implementation issues

There are two main implementation issues that may affect the accuracy of the delay measures provided by our approach: i) matching and synchronization of the S/H circuits used for sampling *Vin* and *Vout*, and ii) generation of random sampling pulses independent from input waveform Vin.
The first problem was addressed by manually placing input and output S/H circuits at the same distance from the DUT and by manually routing their input and control signals. A validation test structure was implemented with the two S/H circuits directly connected to the same signal (without any DUT in between) to provide a null propagation delay. All experiments performed on the validation test chip returned null delay measures. In other words, counters 10 and 01 were stuck at 0, as expected, meaning that the mismatch between input and output samplers does not impair the accuracy of the approach.
Sampling pulse generation was implemented using a variable-period ring oscillator controlled by a pseudo-random signal provided by a maximum-length 8-bit linear feedback shift register (LSFR) [8]. The output of the ring oscillator was used both to control the S/H's and to clock the LFSR, thus updating the period (i.e., the sampling step) at each cycle. The result is a sequence of 255

pseudo-random sampling pulses, repeated periodically. Hence, in principle, the autonomous circuit composed of the ring oscillator and the LFSR provides only 255 independent random samples. However, if the period of the sampling waveform *Tsample* is independent from the period of the input waveform *T*, homologous sampling pulses of different periods may correspond to different sampling points. The maximum number of pseudo-random samples depends on the least common multiple (*LCM*) of *Tsample* and *T*, and is given by

$$N_{ind} = 255\frac{LCM(Tsample,T)}{Tsample} \qquad (6)$$

Our implementation provides two degrees of freedom that can be used to enhance accuracy by increasing the number of independent samples N_{ind}: the period of input signal Clk_{in}, that is directly provided from an external pin, and the supply voltage of the internal ring oscillator (RO), that is independent from the rest of the circuit. The supply voltage of RO can be used not only for tuning *Tsample* at setup time in order to maximize the value of N_{ind} provided by equation (6), but also to dynamically modulate the performance of the RO during operation in order to further increase the number of independent random samples.

Notice that the theoretical number of independent samples N_{ind} provided by equation (6) is not necessarily equal to the numer of samples actually used *N*. In fact, the experiment can terminate with $N<N_{ind}$, or *N* can exceed the number of N_{ind} in order to filter clock jitter ($N>N_{ind}$). From a practical point of view, the total event count to be used in equation (5) is always *N*, while the event count to be used to estimate the standard error due to random sampling is $\min\{N,N_{ind}\}$.

4.2 Choice of time period *T*

In principle, the input time period *T* does not impact the measured propagation delay. In fact, equation (5) expresses *Trr* as the product between *T* and the ratio $N(1,0)/N$, that is inversely proportional to *T*. However, the value of *T* does affect the accuracy of the measure.

With respect to the experiment of sampling and digitizing the input and output signals of a non-inverting cell as described in Section 3, we denote by X a random variable taking value 1 when *Sin*=1 and *Sout*=0, and value 0 otherwise. The expected value of X (denoted by E_X) is the ratio *Trr/T*, while its standard deviation is

$$\sigma_X = \sqrt{\frac{Trr}{T}\left(1-\frac{Trr}{T}\right)} \qquad (7)$$

that can be approximated by

$$\sigma_X \cong \sqrt{\frac{Trr}{T}} \quad \text{for } T >> Trr \qquad (8)$$

By virtue of the theory of random sampling [7], the sample average X_{avg} computed over a sample of *N* independent observations of *X* is an estimator of E_X with standard deviation

$$\sigma = \frac{\sigma_X}{\sqrt{N}} \qquad (9)$$

Notice that E_X is nothing but probability Pr(*Sin*=1,*Sout*=0) introduced in Section 3, while X_{avg} is the relative frequency $N(1,0)/N$ that appears in Equation (5).

Putting equations (5), (8) and (9) together, we can obtain the relation between the standard deviation of the measure of *Trr* and the value of *T*:

$$\sigma_{measTrr} \cong \frac{1}{\sqrt{N}}\sqrt{\frac{Trr}{T}}T = \sqrt{\frac{Trr \cdot T}{N}} \qquad (10)$$

Hence, choosing a time period *T* much larger that the propagation delay of interest leads to a standard error that grows with the square root of *T*.

The best choice for period *T* is the minimum value that guarantees that the distance between rising and falling input transitions is safely greater than the propagation delays under measure.

5. Preliminary experimental results

The test chip presented in this paper will be used to monitor propagation delays during process development and to characterize/validate the delay models of a design kit.

We evaluated measure repeatability by computing the standard deviation of repeated measures (for a fixed timing arc of a fixed DUT on a fixed die). Using a time period of 5ns and 100,000 independent samples to measure a propagation delay of the order of 200ps, we obtained a measured standard deviation of about 1ps (i.e., 0.5%), that is lower than the theoretical value of 2ps provided by equation (10). Both the measured and theoretical standard deviations are within the tolerance of the external instrumentation (power supply and pulse generator).

As an example application, we used the test chip to evaluate the effect of intra-wafer process variations. To this purpose we measured the variations across different dies taken from the same wafer of the propagation delays of two DUTs (namely, an inverter, IVP, and a 2-input nand gate, ND2) with fixed operating conditions.

Table 1 reports the results obtained for a corner lot of a 0.18µm process with extreme variations. The standard

deviation reported in the second-last row represents the effect of process variations on circuit performance. Traditional approaches were also implemented for validation and comparison. In particular, we performed SPICE simulations annotated with measurements based on ring oscillators built of the same cells. The reference values obtained for the four timing arcs are reported in the last row of Table 1. The relative difference between the reference values and the average of the measured propagation delays if within 6%.

6. Conclusions

We have presented a new approach to on-chip delay measures based on random sampling. The proposed approach is independent of the nature of the DUT, making it suitable for measuring the propagation delay of any combinational path across a single cell, a complex combinational unit, or a single interconnect. The distinguishing feature of our technique is the capability of measuring single-cell delays, thus enabling the characterization of corner cases, without using high-performance test equipments. Moreover, the compact implementation enables the integration of many DUTs on the same die.

A prototype test chip with an array of 32 DUTs has been realized in a 0.18µm embedded NVM CMOS technology. Preliminary experimental results have been presented, demonstrating the feasibility and the practical applicability of the proposed technique.

Acknowledgements

This work has been partially supported by Italian National funding, under the project MEDEA + T101 "Technodat".

Die No.	IVP Trf (ps)	IVP Tfr (ps)	ND2 Trf (ps)	ND2 Tfr (ps)
1	142	202	350	454
2	168	204	372	444
3	154	226	356	475
4	158	200	341	420
5	177	189	373	424
6	153	229	365	423
7	160	204	369	443
8	166	223	390	421
9	144	212	402	440
10	162	213	362	450
11	155	200	369	440
12	167	215	380	436
13	152	219	394	413
14	152	205	380	415
15	158	218	386	455
16	171	224	367	423
17	161	218	387	474
18	167	217	351	456
19	177	191	408	434
20	169	210	386	419
21	151	201	382	455
22	172	221	338	421
23	160	188	371	442
24	157	207	383	431
25	169	213	381	423
26	150	228	381	472
Average	**160**	**210**	**375**	**438**
σ	**9.2**	**11.5**	**17.3**	**18.1**
Ref.	**162**	**197**	**330**	**432**

Table 1. Measurement results obtained for two library cells (namely, an inverter and a 2-input nand gate) on 26 different dice taken from the same wafer. Data refer to an output load of 60fF, and an input rise time of 0.1ns.

References

[1] O. Coudert, "Timing and design closure in physical design flows," in *Proc. of ISQED*, pp. 511-516, 2002.

[2] J. A. Davis, et al. "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century," in *Proceedings of the IEEE*, vol. 89, no. 3, pp. 305- 324, 2001.

[3] B. E. Stine, E. Chang, D. S. Boning, and J. E. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," in *IEEE Trans. on Semiconduct. Manufact*., vol. 10, pp.24-41, 1997.

[4] W.H. Kao, Chi-Yuan Lo, M. Basel, and R. Singh, "Parasitic extraction: current state of the art and future trends," in *Proceedings of the IEEE*, vol. 89, no. 5, pp. 729-739, 2001.

[5] N. Abaskharoun, et al., "Circuits for on-chip sub-nanosecond signal capture and characterization," in *Proc. of IEEE Conf. on Custom Integrated Circuits*, pp. 251 -254, 2001.

[6] Y.Arai, M.Ikeno, "A time digitizer Cmos Gate Array with a 250ps Time Resolution" in *IEEE J. Solid-State Circuits*, vol.31, no.2, 1996.

[7] S. K. Thompson, *Sampling, 2nd edition*, Wiley, 2002.

[8] K. S. Trivedi, *Probability and Statistics with Reliability, Queueing, and Computer Science Applications, 2nd edition,* Wiley, 2001.