Leveraging 3D-IC for On-chip Timing Uncertainty Measurements

Randy Widialaksono, Wenxu Zhao, W. Rhett Davis, Paul Franzon
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, North Carolina, USA

***Abstract*—Modern high-performance designs require accurate on-chip timing uncertainty measurements for post-silicon validation of high speed interfaces and clock distribution networks. On-chip timing measurements capabilities must keep up with growing design complexity and process variations to meet competitive product time-to-market. However, enhancing silicon debug capabilities cannot simply be met by proliferating on-chip structures, since the overhead would be prohibitively expensive to deploy. We propose moving on-chip debug and validation structures onto a separate die which would be stacked onto the product die using three-dimensional integration (3D-IC). This paper focuses on achieving observability at clock sinks which are critical for understanding on-chip timing uncertainty. We present a circuit implementation and design flow which realizes high volume on-chip timing measurements for a 2D product die.**

I. Introduction

Modern high-performance designs require accurate on-chip timing uncertainty measurements for post-silicon validation of high speed interfaces and clock distribution networks. These measurements are facilitated by on-chip timing sensors, which incur area, routing, and power overhead. With increasing design complexity and process variations, post-silicon validation and debug capabilities must keep up accordingly to meet competitive product time-to-market. However, enhancing post-silicon validation and debug capabilities cannot simply be met by proliferating on-chip structures, since the overhead would be prohibitively expensive. To enhance validation and debug capabilities without recurring overhead in the high-volume product, we propose moving these validation structures onto a separate die which would be stacked onto the product die using three-dimensional integration (3D-IC). The cost of die stacking would be justified by reducing the product die area or by accelerating silicon debug and validation for faster time-to-market.

This paper focuses on achieving high observability of timing at clock sinks which are critical for timing verification, as illustrated in Fig. 1. Clock sinks are probed for observing the effect of multiple variation sources on the clock distribution network. With high volume on-chip timing sensors, chip designers would have better understanding of on-chip timing uncertainty. While the target design under validation in this paper is a 2D design, the concept could be expanded for validating 3D-IC products.

3D integration also offers stacking of dies that are fabricated with different process technologies (heterogeneous die stacking). The chip designer could take advantage of this feature by implementing the validation die on a more mature process than the product die. To best implement this strategy, the time measurement resolution should be finer than an individual gate delay.

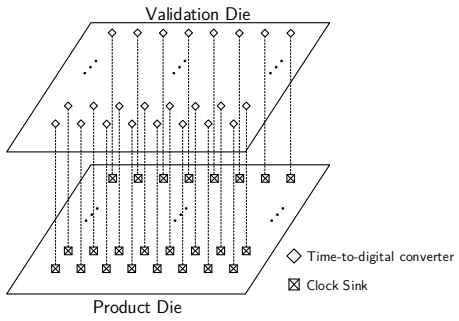


Fig. 1. High volume on-chip timing measurements with 3D-IC

The Vernier time-to-digital converter (TDC) is one of several circuit architectures which achieve sub-gate delay resolution [1]. We expand previous work on timing uncertainty measurements [2] by deploying the Vernier TDC structure with modifications to support measurements in 3D. The snap-on, plug-and-play feature of 3D stacking has been previously proposed for stackable processor design [3], fault-tolerance improvement [4] and software introspection [5].

In this paper we propose a methodology for enhancing on-chip timing measurements through 3D integration. Section 2 describes the proposed methodology and its advantages. In Section 3 we describe our circuit implementation and its operation. Section 4 presents the physical implementation and the design flow.

II. Measurement in 3D

A. Advantages

1) Enhanced On-chip Timing Observability: Placing more on-chip timing sensors would aid post-silicon validation by enhancing the amount of timing information visible to the validation engineer.

2) Reduced Area Overhead: By moving debug structures onto its own die, it will no longer be embedded in the high-volume product die. This concept generally applies to on-chip

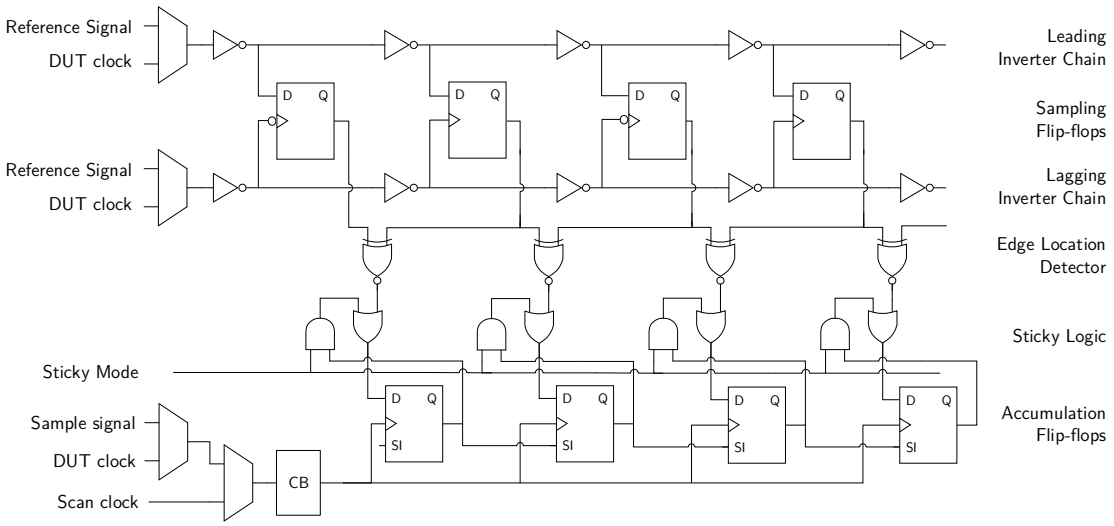


Fig. 2. Proposed Vernier TDC architecture

debug and validation structures.

3) Noise Isolation: By moving the measurement circuits onto its own die, its power supply is better isolated from power supply excursions and crosstalk noise induced by the DUT. This isolation could be further improved by adding adequate number of decoupling capacitors on the validation die.

B. Measurement Methodology

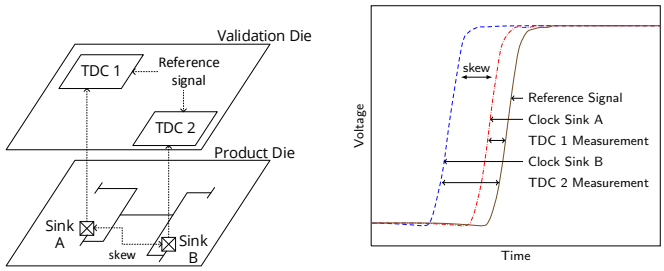
The methodology consists of four major components: time-to-digital converters (TDCs), the clock sinks of interest, die-to-die vias, and a reference edge signal.

1) TDC Requirements: Time-to-digital converters (TDCs) facilitate measurement of on-chip timing uncertainty. There are several possible implementations for a TDC, hence we impose the following requirements for choosing a suitable TDC implementation to the methodology:

- Sub-gate delay measurement resolution: since the process technology of the validation die may differ with the process technology of the product die, the resolution should be independent from the minimum gate delay of a certain process. Hence chip designers could opt to fabricate the validation die in a more mature, reliable process with less design rule checks, less process variations and more accurate models.
- Low power consumption: in order to minimize thermal and power noise that could affect DUT operation.
- Area overhead: the area overhead of the timing sensors should be sufficiently small enough for the desired number of probe points.

Each TDC is connected to a clock sink located in the product die through a (face-to-face) F2F microbump. A reference edge signal is connected to every TDC in the validation die to enable indirect skew measurement.

2) Clock Sinks: The clock sink selection process is left to the chip designer's own discretion. For instance, the designer could select sinks based on its insertion delay relative to the median or prioritize on high switching activity regions.



(a) Reference signal (b) Timing diagram

Fig. 3. Measurement Scheme

3) Inter-die Vias: Ideally the inter-die vias should have high yield and negligible or uniform parasitic delay. If the inter-die vias have large variations in parasitic delay, the deviation would skew measurement results, hence additional processing is needed to characterize and de-embed this component from the sensor readings.

4) Reference Signal: The reference signal on the validation die enables indirect skew measurement as shown in Fig. 3. Each TDC measures the time between the reference signal edge with the clock sink under measurement. Ideally there should be no skew for this reference signal, since it is used as a time reference for every TDC. Constructing a grid/mesh for the reference signal would be a feasible method to implement, considering the relatively ample routing and area resources on the validation die.

If the reference signal tree has large skew, the skew component must be de-embedded from the measurement results through calibration. Otherwise, direct comparison is not possible between TDCs with large reference signal skew. The skew of the reference edge signal is calibrated by setting the signal as inputs to the TDC. When this skew is zero, the readings of the sensors would be identical. Else, the difference denotes the skew amount.

III. Circuit Implementation

A Vernier delay line based TDC is implemented as shown in Fig. 2. The Vernier delay line is known for its sub-gate delay time resolution and robustness against process, voltage, and temperature (PVT) variations [1].

The sampling flip-flops function as an early-late detector between the two inverter chains. The edge location detection logic is mainly used for multi-cycle jitter measurements. The accumulation flip-flop chain captures and stores measurement results until scan-out.

A. Clock Skew Measurement

As illustrated in Fig. 3, clock skew is measured by comparing different DUT clocks to a reference signal. The DUT clock and the reference signal are sent down the leading and lagging inverter chain respectively. The sampling flip-flop chain indicates how long the leading edge have propagated before the lagging edge arrives. Afterwards, the sampled signals are processed by the edge detection logic and latched into the accumulation flip-flops when triggered by the Sample signal. This Sample signal could be generated by delaying the reference signal for the edge detection logic propagation delay, or independently. The time between the reference edge and the DUT clock edge is indicated by multiplying the number of consecutive zeros with the Vernier delay line resolution. The skew between two clock sinks is finally calculated by comparing the outputs of its TDCs.

B. Clock Jitter Measurement

Jitter measurement is launched by setting the circuit to sticky mode and feeding the DUT clock into three inputs: the leading inverter chain, the lagging inverter chain and the clock buffer for the accumulation flip-flop chain. The changing location of edges indicates jitter. By asserting the sticky mode signal, the worst case jitter is recorded by the accumulation flip-flop chain. Jitter measurement is concluded by switching the circuit into scan-out mode.

C. Area & Power Consumption

An implementation of the TDC architecture which contains 150 inverters in each delay line, sampling and accumulation flip-flops, and processing logic, as shown in Fig. 2, consumes 16,931 $\mu m2$ and 4284,72 $\mu m2$ of area in a commercial 130 nm and 65 nm process technology respectively. When laid out with a 1:1 aspect ratio, the TDC pitch would be approximately 130 μm and 65 μm in 130 nm and 65 nm process technology respectively. A validation die with 5 mm x 5 mm dimensions could contain at most 1,444 and 5,929 TDCs in 130 nm and 65 nm process technology respectively.

Power and energy consumption were measured to predict thermal and power impact by using Cadence® Spectre® for transistor level simulation. Power consumption for jitter measurements is 2.11339 mW in 130 nm technology, assuming input clock frequency of 2 GHz. Energy consumption for a one-shot skew measurement is 1.057 pJ in 130 nm technology.

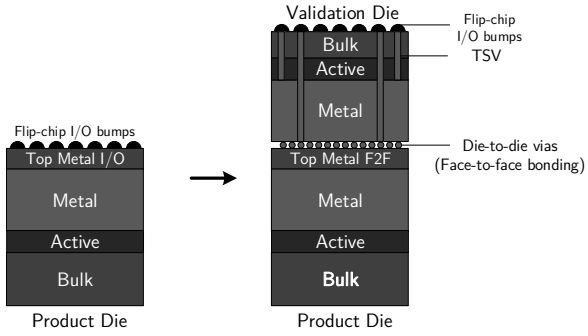


Fig. 4. Cross section view of proposed physical implementation

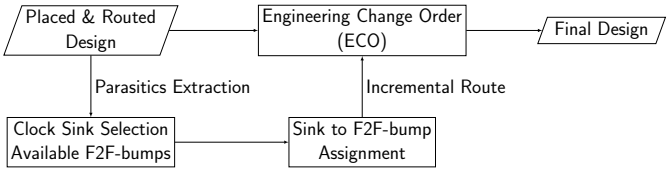


Fig. 5. Design Flow

IV. Physical Implementation

A. Die Stack Configuration

Between the two die stacking techniques: Face-to-Face (F2F) and Face-to-Back (F2B), the F2F bonding would be the better option because F2F bumps have much lower parasitic and smaller pitch than Through-Silicon-Vias (TSVs). F2F bondpoints have negligible delay parasitics, hence allowing on-chip timing measurements without imposing additional measurement offsets. F2F bondpoints commonly have 10-25 μm [6] pitch, hence providing fast, high throughput interface between the product and validation die. By using the F2F interface, the maximum number of on-chip sensors would not be constrained by the inter-tier bump pitch, but by the size of the TDC.

For tier assignment, the validation die would be designated as the tier with its substrate thinned for I/O connections. This assignment is to avoid modifications to the design on the product die. The product die would then have its I/O connections routed through the F2F bondpoints, along with additional signals of interest for debug and validation. The die stack configuration is shown in Fig. 4.

B. Design Flow

The physical design challenges include selecting which clock sinks to probe and assigning those selected sinks to available bondpoints/bumps. The design flow as shown in Fig. 5 starts with a placed and routed design with its parasitic extracted. This is to ensure we obtain clock sink insertion delay analysis that are as accurate as possible.

After the initial place and route step, the available bondpoints are those that are not reserved for routing the product's I/O signals in both configurations shown in Fig. 4. The routing between the tapped clock sinks and the inter-tier bumps

become stubs in the final product, with low impact on power and area as described in [5].

C. Inter-tier Bump Assignment

The bump assignment step considers the interconnect delay between the clock sink its assigned F2F bump. If the parasitic delay matches between two TDCs, hence the difference of both outputs would indicate actual skew. Otherwise, the parasitic delay must be de-embedded from the measurement readings. One way to avoid characterizing and de-embedding parasitic delay is to match the wire delay between a selected clock sink and its assigned bump.

Matching wire delays could be achieved through custom routing or wire delay matching. Custom routing allows skew measurement between any clock sink regardless of their proximity to a F2F bump, as long as their wire delay are matched. Without custom routing, we approximate wire delay by calculating the Manhattan/rectilinear distance between a selected sink with a candidate bump, and then verifying with parasitic extraction.

In a scenario where not all selected sinks could be assigned to a bump with matching delays, the designer could prioritize selected sinks into groups based on wire delay or distance to assigned bump.

The bump assignment algorithm is shown in Fig. 6. First a *k-d* tree [7] is built from available bump locations for fast nearest available bump search. Then each sink group is assigned to available bumps sequentially based on its priority order.

The $AssignSinkGroup$ function assigns each clock sink to a bump. First the function searches the nearest available bumps up to an amount specified by the designer. The designer specifies the maximum distance between a selected sink with its assigned bump. This distance should be kept as small as possible, and increased gradually when constrained by the availability of nearby bumps. The designer also specifies the accepted distance range, namely e. The algorithm continues to build the list of nearest bump until the desired distance is reached, with exponentially increasing amount of bump added on every search for speed-up. If a candidate available bump is within the desired range, it is assigned to the clock sink and removed from the *k-d* tree of available bumps. The function returns a list of successfully assigned clock sinks. Finally the sink-to-bump assignments are incrementally routed in the Engineering Change Order (ECO) step.

Input: sg_l: List of sink groups
Input: b_l: List of available bump locations
Input: d: Distance from sink to bump
Input: e: Distance tolerance
Input: n: Obtain n nearest bump per iteration
Ensure: Available bumps (b_l) $>$ tapped sinks
Output: a: List of sink-bump assignments

function ASSIGNALLGROUPS(sg_l, b_l, d, e, n)
  $kd \leftarrow kdtree(b\_l)$
  for all $sg \in sg\_l$ **do**
    append(a, AssignSinkGroup(sg, kd, d, e, n))
  end for
  return a
end function

function ASSIGNSINKGROUP(sg, kd, n, d, e)
  for all $s \in sg$ **do**
    $bl \leftarrow$ sort_ascend (n_nearest (s, kd, n))
    while tail(bl)$< d$ **do**
      $n \leftarrow n * 2$
      $bl \leftarrow$ sort_ascend (n_nearest (s, kd, n))
    end while ▷ query n-nearest neighbor until list contains bump that reaches distance
    for all $b \in bl$ **do**
      if manhattan_distance(s, b) $= d \pm e$ **then**
        append($m, (s, b)$) ▷ m: intermediate list
        remove(kd, b)
        break
      end if
    end for
  end for
  return m
end function

Fig. 6. F2F Bump Assignment Algorithm

V. CONCLUSION

In this paper, we proposed a methodology for achieving high-volume on-chip timing uncertainty measurements without incurring silicon area overhead on the product die, through 3D integration. We designed a timing measurement macro based on the Vernier delay line and the Skitter circuit [2]. We proposed a physical design flow and a bump assignment algorithm for validating a 2D product die.

REFERENCES

[1] P. Dudek, S. Szczepanski, and J. Hatfield, "A high-resolution cmos time-to-digital converter utilizing a vernier delay line," *Solid-State Circuits, IEEE Journal of*, vol. 35, no. 2, pp. 240–247, Feb 2000.

[2] R. Franch, P. Restle, N. James, W. Huott, J. Friedrich, R. Dixon, S. Weitzel, K. Van Goor, and G. Salem, "On-chip timing uncertainty measurements on ibm microprocessors," in *Test Conference, 2007. ITC 2007. IEEE International*. IEEE, 2007, pp. 1–7.

[3] E. Rotenberg, B. Dwiel, E. Forbes, Z. Zhang, R. Widialaksono, R. Basu Roy Chowdhury, N. Tshibangu, S. Lipa, W. Davis, and P. Franzon, "Rationale for a 3d heterogeneous multi-core processor," in *Computer Design (ICCD), 2013 IEEE 31st International Conference on*, Oct 2013, pp. 154–168.

[4] N. Madan and R. Balasubramonian, "Leveraging 3d technology for improved reliability," in *40th Annual IEEE/ACM International Symposium on Microarchitecture, 2007. MICRO 2007*, Dec. 2007, pp. 223–235.

[5] S. Mysore, B. Agrawal, N. Srivastava, S.-C. Lin, K. Banerjee, and T. Sherwood, "3d integration for introspection," *Micro, IEEE*, vol. 27, no. 1, pp. 77–83, Jan 2007.

[6] P. Enquist, G. Fountain, C. Petteway, A. Hollingsworth, and H. Grady, "Low cost of ownership scalable copper direct bond interconnect 3d ic technology for three dimensional integrated circuit applications," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*. IEEE, 2009, pp. 1–6.

[7] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, Sep 1975.