Low-Overhead Design Technique for Calibration of Maximum Frequency at Multiple Operating Points

Somnath Paul*, Sivasubramaniam Krishnamurthy*, Hamid Mahmoodi* and Swapup Bhunia*

*Electrical Engineering and Computer Science Department, Case Western Reserve University, Cleveland, OH.
*Electrical and Computer Engineering Department, San Francisco State University, CA.

Abstract— Determination of maximum operating frequencies (\(F_{\text{max}}\)) during manufacturing test at different operating voltages is required to: (a) to ensure that, for a Dynamic Voltage and Frequency Scaling (DVFS) system, the adaptation hardware actually applies the correct operating frequency corresponding to a scaled supply and (b) to sort chips in different voltage-frequency (V-F) bins, so that chips at different bins can be used for different applications. Existing speed binning approach requires extensive delay testing at all operating points with all possible frequencies, which increases test cost and test time significantly. In this paper, we propose a low-overhead solution for characterizing \(F_{\text{max}}\) of a circuit at different operating voltages that can eliminate the complex and expensive \(F_{\text{max}}\) calibration at multiple voltage points. The basic idea is to choose a small set of representative paths in a circuit based on their voltage sensitivity and dynamically configuring them into ring oscillator to compute \(F_{\text{max}}\). The proposed calibration mechanism is all-digital, robust to process variations, reasonably accurate (average 2.8% error) and incorporates minimal hardware overhead (average 1.7% delay, 3.5% area and 0.28% power overhead).

Index Terms: Dynamic Voltage and Frequency Scaling, Frequency calibration, Voltage Sensitivity, Ring Oscillator.

I. INTRODUCTION

With continued technology scaling, process parameter variations (both systematic and random) and other design marginalities grow larger, resulting in increasing fluctuations in the delay of a logic circuit. High-performance circuits such as microprocessors, graphics processors or digital signal processors suffer significantly from delay/frequency fluctuation, which is typically estimated to be about 30% in a microprocessor [7]. This wide variation in frequency has introduced the concept of frequency or speed binning [10]. It has become essential to perform speed binning of high-performance chips not only to earn extra profit for the higher performance chips but also to salvage the slower but non-faulty chips in a possible “go-no-go” situation.

Speed-binning requires the calibration of \(F_{\text{max}}\) at different operating conditions such as supply voltage, temperature etc. In the simplest scenario, it is desired to determine \(F_{\text{max}}\) corresponding to a given operating voltage under worst-case temperature condition. The process is expensive in terms of both test application time and complexity of the test hardware since it requires testing at multiple frequencies for a given supply voltage. The situation becomes worse when it is required to calibrate \(F_{\text{max}}\) at multiple operating voltages. Determination of maximum operating frequencies during manufacturing test at different operating voltages is required: (a) to ensure that, for a Dynamic Voltage and Frequency Scaling (DVFS) system (e.g. one that adjusts its operating voltage/frequency dynamically to reduce power or operate under a power-temperature envelope [6]), the adaptation hardware actually applies the correct operating frequency corresponding to a scaled supply and (b) to sort chips in different voltage-frequency (V-F) bins, so that chips at different bins can be used for different applications. Using Monte-Carlo simulations of some benchmark circuits, we have observed that frequency vs. voltage relationship not only changes from chip to chip but changes unpredictably at different voltage points for the same chip as well. Thus a static design-time calibration does not work well unless it is very conservative.

Given the complexity and cost of speed binning at just one voltage, it can be easily envisaged that determination of \(F_{\text{max}}\) at multiple voltage points will be a very expensive process. Since we need post-silicon calibration of \(F_{\text{max}}\) at different operating voltages, it is important to develop techniques that aid the speed binning process based on structural testing to reduce its cost. Earlier it has been demonstrated that speed binning using structural delay testing correlates very well with functional speed binning [10]. There are two major techniques to improve the efficiency of speed binning. The first approach is based on creating a critical-path replica to estimate the delay of critical path after manufacturing [8]. However, the replica cannot reliably represent the delay of the actual critical path due to local within-die delay variations arising from effects like random dopant fluctuations. Hence, in order to measure the frequency shift accurately, we need to consider the actual timing paths in the circuit. In the second approach, on-chip delay measurement hardware is used to estimate the delay of a critical path after manufacturing [11]. Although this approach targets actual critical path of a circuit, it requires a high-speed clock or on-chip reference voltages to estimate path delay, both of which impose major design challenges. Moreover, to achieve good confidence of \(F_{\text{max}}\) estimation, we need to insert delay sensors at large number of timing paths, which incurs large design overhead.

In this paper, we propose a low-overhead and robust mechanism for \(F_{\text{max}}\) against \(V_{\text{DD}}\) calibration that takes into account actual timing paths. The proposed method requires conventional calibration at the nominal voltage (delay testing at multiple frequencies) and based on these delay results, it computes \(F_{\text{max}}\) for other voltage points efficiently, by applying very few delay test vectors to estimate \(F_{\text{max}}\) with reasonable accuracy. The method works in three steps. 1) Few representative timing paths which are critical and also highly voltage sensitive are selected for delay calibration. Voltage sensitivity is measured by the vulnerability of path delay to a voltage shift. 2) During manufacturing test, each selected path is dynamically configured into ring oscillator by the application of a control signal and its frequency is determined using a simple frequency counter. The oscillation frequencies of the selected paths are then used to estimate their voltage sensitivity during the manufacturing test. 3) The critical delay of the circuit at a particular operating voltage is then estimated on the basis of the most critical delay under nominal operating conditions and the maximum voltage sensitivity at the nearest calibration voltage. The paper makes the following contributions:

1. It addresses two important test considerations. a) To select few timing paths for voltage-frequency calibration from the exponentially large number of timing paths existing in the design. b) To measure the delays for the timing paths by a method that incurs low design overhead and is tolerant to process variation.

2. It proposes selection of timing paths based on their voltage sensitivity for \(F_{\text{max}}\) calibration. It shows that the relative sensitivity of timing paths remain unchanged with process variations, although their absolute value may change.

3. It proposes dynamically configuring the selected paths into ring oscillator and computation of \(F_{\text{max}}\) at a particular voltage point.
from the voltage sensitivities of those paths and the critical delay measured at nominal operating conditions.

II. SELECTION OF TIMING PATHS BASED ON VOLTAGE SENSITIVITY

In this section, we propose the selection of the critical paths on the basis of their voltage sensitivity for \( F_{\text{max}} \) calibration. The delay through a logic gate changes with variation in supply voltage. Such a variation is non-linear and is also a function of the process and device parameters [5]. In order to have an estimate of the delay variation of the logic gate with supply voltage we define a metric called \( \text{voltage sensitivity} \), which will be able to capture the change in delay with change in supply voltage. An analytical expression for voltage sensitivity can be derived using the Newton-Sakurai Alpha Power model as described in [9]. According to this model, for voltage sensitivity can be derived using the Newton-Sakurai model described in [9].

The basis of their voltage sensitivity for the charge and discharge time of a single transistor output load model, as described in [9]. According to this model, for voltage sensitivity can be derived using the Newton-Sakurai model described in [9].

\[
\alpha = \frac{C_{\text{ox}}}{L} \cdot \frac{V_{\text{DD}}}{C_L} \cdot K_{N,P}(\theta) \left[ V_{\text{DD}} - V_{T_{\text{NO}}P}(\theta) \right] \partial V_{\text{DD}} - 2C_{N,P} \partial \theta
\]

Equation (1)

Parameters \( K_{\theta}(\theta) \) and \( V_{T_{\theta}}(\theta) \) depend on operating temperature \( \theta \) as

\[
K(\theta) = K_{\theta_0,\theta}(\theta_{\text{nom}}) \frac{\partial \theta}{\partial \theta}
\]

\[
V_{T_{\theta}}(\theta) = V_{T_{\theta_0}}(\theta_{\text{nom}}) - \delta(\theta - \theta_{\text{nom}})
\]

The delay sensitivity \( S(V_{\text{DD}}, \theta) \) for low-high transition or high-low transition is defined as:

\[
S(V_{\text{DD}}, \theta) = \frac{\partial \theta}{\partial \theta} \frac{\partial \theta}{\partial V_{\text{DD}}}
\]

For a change in temperature only at constant supply voltage, the delay sensitivity can be approximated by the first order temperature sensitivity given by,

\[
S(\theta) = \frac{\partial \theta}{\partial \theta} \frac{\partial \theta}{\partial V_{\text{DD}}}
\]

Similarly for the case, when the circuit is operating at constant temperature under varying supply voltage, we may define a first order voltage sensitivity given by,

\[
S(V_{\text{DD}}) = \frac{\partial \theta}{\partial \theta} \frac{\partial \theta}{\partial V_{\text{DD}}}
\]

The definition of delay sensitivity as given by equation (4) can be used to accurately model the delay in a change in the operating conditions such as supply voltage and temperature. This can be easily seen from the following Taylor Series expansion which models the change in delay due to change in voltage (\( \Delta V_{\text{DD}} \)) and temperature (\( \Delta \theta \)).

From equation (7) we see that the voltage and temperature sensitivities given by \( S(V_{\text{DD}}) \) and \( S(\theta_{\text{TDD}}) \) depends on the process parameter \( \theta_{\text{LNO}}, \theta_{\text{NNO}}, \delta_{\text{LNO}}, \delta_{\text{NNO}}, V_{\text{TNO},P}(\theta_{\text{nom}}) \) (Threshold voltage at nominal operating temperature \( \theta_{\text{nom}} \)), the operating temperature \( \theta \) and the operating voltage \( V_{\text{DD}} \). Hence the switching delay and the delay sensitivity through a transistor vary with voltage and temperature. The idea can be extended to complex logic gates and thus to timing paths whose delay will vary with voltage and temperature. Calibration of the maximum allowable frequency against different supply voltages requires a prior knowledge of the voltage sensitivity of the different logic paths. Thus, in order to estimate the absolute value of the voltage sensitivity for a logic path, the sizes of the individual gates, the parasitics, the process and device parameters, operating temperature and the operating voltage should be taken into consideration. As seen from equation (7) the measured value of voltage sensitivity is able to account for non-linear variation in delay with supply voltage at different temperature corners. In order

\[
S(V_{\text{DD}}) = \frac{\partial \theta}{\partial \theta} \frac{\partial \theta}{\partial V_{\text{DD}}}
\]

\[
S(\theta_{\text{TDD}}) = \frac{\partial \theta}{\partial \theta} \frac{\partial \theta}{\partial V_{\text{DD}}}
\]

Figure 1: a) Delay variation of a critical path in a benchmark circuit (ISCAS85 c880) with voltage; b) corresponding variation in voltage sensitivity.

Figure 2: Variation in Voltage Sensitivity for two paths at different process corners.

III. SELECTION OF SENSITIVE PATHS

The choice for paths having maximum value of voltage sensitivity at the calibrating voltages is performed at the time of calibration. The delay sensitivity at the calibrating voltages is performed at the time of calibration. The delay sensitivity at the calibrating voltages is performed at the time of calibration. The delay sensitivity at the calibrating voltages is performed at the time of calibration.
design. It is not sufficient to choose a set of paths only on the basis of their criticality at nominal supply voltage for estimation of critical delay at different voltage corners. The reason is: it is likely that the most critical path of the design is not the most voltage sensitive path. A path which is less or equally critical at a higher voltage value can have a higher path delay at lower voltage due to its greater voltage sensitivity. Hence, the problem can be defined as identifying the most voltage sensitive paths of the circuit at the time of the design, followed by the measurement of their voltage sensitivities at the different calibration steps. The measured values of voltage sensitivities are used in conjunction with the critical delay at nominal operating conditions for delay estimation at other voltage corners. The selection of most voltage sensitive paths is done from a set of timing paths that are within a specified margin of the critical delay at nominal conditions.

Different regions of the chip normally operate at different temperature due to different local activity and integration density at different regions of the die. This creates local heating giving rise to what is known as ‘hotspots’. Typically, the most critical paths are found near the hotspot areas. The proposed delay estimation framework works even in the case of timing paths subject to different temperature values. The reason is, voltage sensitivity being a function of operating temperature, the measured values of voltage sensitivity is able to account for the changes in operating temperature. In the proposed method, the most voltage sensitive path at temperature θ and nominal supply voltage VDD be denoted as Dcri(VDD, θ). Let AVDD be the change in voltage and the new value of critical delay be given by Dcri(VDD + AVDD, θ). Let us assume that the voltage sensitivity of the most voltage sensitive path at a temperature θ and voltage VDD is known to us and let us denote it by S(VDD, θ). Then using the value of this sensitivity the delay Dcri(VDD + AVDD, θ) can be estimated by equation (8):

\[ D_{\text{cri}}(VDD + AVDD, \theta) = D_{\text{cri}}(VDD, \theta) + S(VDD, \theta) \times AVDD \]  

(8)

where the voltage sensitivity is given by the equation (9) and D(VDD, θ) denotes the delay of the most voltage sensitive path at voltage VDD and temperature θ.

a) Estimation of delay at different supply voltages The voltage sensitivity metric as defined in equation (7) can be used to find out the critical delay for a given design at a voltage point different from the nominal operating point. Let us assume that the critical delay at the temperature θ and nominal supply voltage VDD be denoted as Dcri(VDD, θ). Let AVDD be the change in voltage and the new value of critical delay be given by Dcri(VDD + AVDD, θ). Let us assume that the voltage sensitivity of the most voltage sensitive path at a temperature θ and voltage VDD is known to us and let us denote it by S(VDD, θ). Then using the value of this sensitivity the delay Dcri(VDD + AVDD, θ) can be estimated by equation (8):

b) Methodology for selection of paths

We define a margin for selection of the paths whose voltage sensitivity is to be noted. From our observations with standard benchmark circuits, voltage sensitivity measurements for paths that lie within 10% of the most critical delay at nominal conditions allow for a good estimate of critical delay at other voltage points. The methodology adapted to estimate the delay calibration can be thus summarized as follows:

1) All the paths having a delay within 10% of the most critical delay at the nominal supply voltage and temperature are chosen at the time of the design. The voltage sensitivities for these paths at the different calibration voltage points are determined using circuit simulations. The paths corresponding to the maximum value voltage sensitivity at the different calibration steps.

2) Suitable design changes are made to the paths corresponding to the maximum voltage sensitivity at different calibrating steps so that they may be dynamically configured as ring oscillators for online measurement of their voltage sensitivity. The maximum voltage sensitivity value is then selected from the set of timing paths configured as ring oscillators for online estimation of critical delay. Voltage sensitivity values measured online may be different from the values obtained at the time of the design due to the variation in operating temperature and process parameters. However, the selection of the timing path with the maximum value of the voltage sensitivity ensures the selection of the correct voltage sensitivity value.

3) The maximum value of voltage sensitivity calculated at each calibrating step is then used to estimate the critical delay using equation (8) at different voltage points.

IV. DYNAMIC RECONFIGURATION OF SENSITIVE PATHS INTO RING OSCILLATOR

We propose an all-digital, robust delay measurement technique that involves reconfiguring the timing paths dynamically into a ring oscillator during the calibration step. The oscillation period for the ring oscillator will then serve as an estimate of the critical path delay. A set of multiplexers needs to be integrated at the inputs and outputs of the circuit to configure multiple timing paths into ring oscillators and measure their oscillation frequency one at a time.

To create a ring, outputs of the critical paths are routed back to the corresponding input terminal. In case a particular timing path has an even number of inverting gates, an extra inverter is introduced at the beginning of the path to make it oscillate. A low frequency reference clock along with a frequency counter is integrated to count the number of oscillation pulses within the duration of a reference pulse. It is to be noted that we need to use control signals to activate the select lines of the input and output multiplexers such that critical paths are sensitized for oscillation one after another. Moreover, appropriate control signals are required to be applied to the non-critical inputs to sensitize the critical path. The control signals can be scanned-in as part of the test vector. The following procedure is proposed for sensitivity computation:

1. For a given calibration step, load the right vector in the scan chain (the vector that sensitizes the desired path).
2. Configure the multiplexers such that the desired path oscillates.
3. Measure the frequency (cycle time) of the desired path from the counter output.
4. Calculate the value of voltage sensitivity using the delay at present voltage and the delay at previous voltage.
5. Repeat the calculation for voltage sensitivity for all timing paths configured as ring oscillators.
6. Choose the maximum value of voltage sensitivity for the given calibration step.
7. Estimate the critical delay for other voltage corners on the basis of the maximum value of measured voltage sensitivity and the critical delay at the nominal operating conditions.

For the ring oscillator to oscillate, the state of other inputs needs to be set such that all the logic gates in the critical path pass the signal transition on their input. The output of the frequency measurement block is a digitized delay measurement of the oscillation cycle time. Since there are two possible signal transitions on any node, there are two delays associated with the critical path, rise delay (t_r) and fall delay (t_f). t_r and t_f are not necessarily equal and either one could be the critical delay.
In the example shown in Figure 3, the voltage sensitive path of the circuit chosen at the time of design originates from input B and ends at output O2 of the combinational circuit. The oscillation path has to also include the slave latch of the input flip-flop and the master latch of the output flip-flop in order to include the flip-flop delay in the oscillation cycle time. Hence, the feedback is provided from the output of the master latch of the output flip-flop to the input of the slave latch of the input flip-flop. The oscillation path is formed of the shaded logic gates in Figure 3. Signal TEST is the configuration signal. When TEST is zero, the circuit is in the normal mode of operation. When TEST is one, the input of the oscillation path is connected to its output forming a ring oscillator.

V. RESULTS

To verify the effectiveness of the proposed technique, we considered a set of large benchmarks from ISCAS89 suite and synthesized them in IBM 0.13µm technology using Synopsys design compiler. For each circuit, we selected 10% of most critical paths using Synopsys static timing analysis tool and compared their voltage sensitivity at seven different voltage points (0.6V, 0.8V, 1V, 1.2V, 1.5V, 1.8V, 2.1V). Partitioned spice netlists were generated for the critical paths and their voltage sensitivities were then computed for each voltage step using the formula presented in Section II. Paths with the maximum value of voltage sensitivity at each voltage step were chosen. Table II lists the total number of unique paths with the maximum value of voltage sensitivity at the different calibration steps for the benchmarks.

### A. \(F_{\text{cal}}\) Estimation Accuracy

The estimation accuracy for the benchmark circuits at three different temperature points is presented in Table I. The estimated delay for the most voltage-sensitive path at a supply voltage \(V\) is given by:

\[
\text{Estimated-Delay} = D_{\text{nom}} + S(V) * (V - V_{\text{nom}}) * D_{\text{nom}}
\]

Hence, the estimation error can be computed as:

\[
\text{Estimation-Error} = \frac{D_{\text{nom}} + S(V) * (V - V_{\text{nom}}) * D_{\text{nom}} - D_{\text{actual}}}{D_{\text{actual}}} * 100%
\]

Here, \(S(V)\) is the sensitivity of the most voltage-sensitive path at the calibration voltage \(V\). \(D_{\text{nom}}\) is the maximum delay at nominal operating voltage and \(D_{\text{actual}}\) is the actual maximum delay at voltage \(V\) obtained from delay simulation of all critical paths using Hspice. Thus, the estimated delay at a particular voltage point is computed conservatively – i.e. it is always overestimated ensuring no functional failure due to underestimation. However, as the accuracy results in Table I suggest, the error due to overestimation is within tolerable limits. The accuracy result in Table I also includes the error due to the ring oscillator-based delay measurement. As shown in Table I, \(F_{\text{cal}}\) estimation error averaged over different supply voltages is very small (between 1 and 4%), although the maximum error is about 8%.

### B. Overhead calculation

Overhead is incurred due to the addition of multiplexers at the inputs of the most voltage sensitive paths, at the output of these paths and use of a frequency counter (as shown in Fig. 3). Depending upon whether there are odd or even number of inverting stages in a path, we may need to add inverter in a path to make the total number of inverting logic to be odd in the ring oscillator configuration. Table II shows the overhead results. It can be observed from the table, that the average overhead is 1.7% in delay, 0.28% in power and 3.5% in die-area. The last column of Table I lists the number of clock cycles required for calibration, which depends on the number of paths considered (Column 2). It is not possible to concurrently measure the voltage sensitivity of all the paths chosen due to the fact that the timing paths may share more than one gate. It requires 50 cycles to compute the delay of a path. Thus for benchmark s1196, the total time required for calibration of the 3 voltage sensitive paths is 150 cycles.

Table I. Estimation Accuracy (0.13µ CMOS; \(V_{\text{nom}}=1.2V\))

<table>
<thead>
<tr>
<th>ISCAS89 Ckt</th>
<th>Temp = 80°C</th>
<th>Temp = 100°C</th>
<th>Temp = 120°C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max%</td>
<td>Avg%</td>
<td>Max%</td>
</tr>
<tr>
<td>s1196</td>
<td>6.920</td>
<td>3.508</td>
<td>5.419</td>
</tr>
<tr>
<td>s1423</td>
<td>7.241</td>
<td>3.825</td>
<td>7.197</td>
</tr>
<tr>
<td>s5378</td>
<td>8.482</td>
<td>3.543</td>
<td>7.066</td>
</tr>
<tr>
<td>s9234</td>
<td>8.391</td>
<td>3.460</td>
<td>7.958</td>
</tr>
<tr>
<td>s13207</td>
<td>2.583</td>
<td>1.448</td>
<td>2.876</td>
</tr>
<tr>
<td>s15850</td>
<td>5.374</td>
<td>2.925</td>
<td>5.282</td>
</tr>
<tr>
<td>s35932</td>
<td>8.493</td>
<td>4.273</td>
<td>5.788</td>
</tr>
</tbody>
</table>

Table II. Overhead results (0.13µ CMOS; 1.2V, 100°C)

<table>
<thead>
<tr>
<th>ISCAS89 Ckt</th>
<th># of paths</th>
<th>% Area overhead</th>
<th>% Delay overhead</th>
<th>% Power overhead</th>
<th># of calibration cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1196</td>
<td>3</td>
<td>9.22%</td>
<td>1.79%</td>
<td>0.72%</td>
<td>150</td>
</tr>
<tr>
<td>s1423</td>
<td>4</td>
<td>6.53%</td>
<td>0.72%</td>
<td>0.49%</td>
<td>200</td>
</tr>
<tr>
<td>s5378</td>
<td>3</td>
<td>2.23%</td>
<td>2.13%</td>
<td>0.13%</td>
<td>150</td>
</tr>
<tr>
<td>s9234</td>
<td>3</td>
<td>3.77%</td>
<td>1.62%</td>
<td>0.4%</td>
<td>150</td>
</tr>
<tr>
<td>s13207</td>
<td>3</td>
<td>0.9%</td>
<td>1.30%</td>
<td>0.06%</td>
<td>150</td>
</tr>
<tr>
<td>s15850</td>
<td>3</td>
<td>0.85%</td>
<td>1.08%</td>
<td>0.05%</td>
<td>150</td>
</tr>
<tr>
<td>s35932</td>
<td>3</td>
<td>1.16%</td>
<td>3.20%</td>
<td>0.14%</td>
<td>150</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

Calibration of maximum operating frequency at different operating points (with respect to voltage) is an expensive process in terms of test time and test cost. In this paper, we propose a low-overhead and efficient design technique for calibration of \(F_{\text{cal}}\) at different voltage points. The proposed method is based on selection of few representative timing paths depending on their voltage sensitivity and then dynamically reconfiguring them as ring oscillators to compute maximum operating frequency of a circuit under varying supply voltage. Simulation results support that the proposed calibration methodology is capable of estimating voltage-induced delay variations in a circuit with reasonable accuracy; is robust with respect to inter and intra-die process variations; and incorporates low design overhead.

REFERENCES