Online Measurement of Timing in Circuits: for Health Monitoring and Dynamic Voltage & Frequency Scaling

Joshua M. Levine, Edward Stott, George A. Constantinides and Peter Y.K. Cheung
Department of Electrical and Electronic Engineering
Imperial College London
London, United Kingdom
Email: {josh.levine05, edward.stott07, g.constantinides, p.cheung}@imperial.ac.uk

***Abstract*—Reliability, power consumption and timing performance are key considerations for the utilisation of field-programmable gate arrays. Online measurement techniques can determine the timing characteristics of an FPGA application while it is operating, and facilitate a range of benefits. Degradation can be monitored by tracking changes in timing performance, while power consumption can be reduced through dynamic voltage scaling (DVS) of the power supply to exploit any spare timing headroom. If higher performance is the objective, dynamic frequency scaling (DFS) can be used to maximise operating frequency. In both cases, online timing measurement of the application circuit is used to exploit favourable operating conditions.**

This work demonstrates a method of online measurement, achieved by sweeping the phase of a secondary clock signal, driving additional shadowing registers strategically added to the application design. The measurement technique and initial voltage and frequency scaling experiments are demonstrated on an Altera Cyclone III FPGA. Timing performance can be measured with a best case resolution of 96ps. The additional circuitry results in minimal overhead in terms of area and performance. Power savings of 23% dynamic and 13% static in an example circuit are achieved through DVS, or performance improvements of 21% through DFS, when compared with operating at nominal core voltage, or timing model F_{MAX}.

I. Introduction

Recent research into VLSI degradation mechanisms suggests that timing performance deterioration could be a major concern in future process technologies [1]. A number of techniques have been proposed for detecting timing faults, errors induced by paths failing to meet timing requirements, in circuits while they are operating online, but it would be advantageous to quantify the circuit's performance. A measurement of timing slack at the end of critical paths achieves this. Reductions in the amount of slack indicate that the circuit is degrading, or ageing, and changes due to temperature, voltage and other fluctuations can be tracked, making timing slack a good measure of health.

In order to achieve reliable operation it is useful to be able to detect impending failure, a good example of a system that can do so is Self-Monitoring, Analysis and Reporting Technology (SMART) [2], which monitors various indicators of the reliability of hard disk drives. SMART is able to advise users to back-up, before a drive failure causes the loss of data. In a similar way, early warning of deterioration in digital circuits could be used to trigger preemptive action and avoid dataloss, downtime or even, in certain applications, risks to property or life.

In this work, a method is proposed that can acquire timing performance data from a circuit, while it is operating, without affecting this operation.

The original contributions are:

- a method of precisely quantifying the timing slack available at a circuit node,
- a calibration technique to remove reliance on timing models and assumptions,
- a method for selecting nodes in a circuit to be monitored, and
- a demonstration that the measurements can be used to improve system performances and reduce power consumption.

II. Related Work

The methods presented in this work build on existing techniques for performing measurement and detecting timing errors in circuits.

A. Dedicated Timing Inference Circuits

An existing method of monitoring timing changes in a device is to use a dedicated measurement circuit known as a tunable replica circuit (TRC) [3]. The circuit is designed to behave in a similar fashion to the critical paths in the application circuit so that its timing can be used to infer timing performance. The TRC does not form part of the design functionality of the system, so it can be measured without affecting operation. The principle depends on changes occurring in a deterministic fashion such that the behaviour of the TRC is representative of changes in the main circuit, this results in the need to used it conservatively.

B. Offline Timing Measurement Circuits

Established timing measurement methods evaluate the delay of circuit paths by detecting timing violations that are induced by a sweep of the clock frequency. The induced

errors are detected through additional circuitry [4], [5] or changes in output distribution [6]. These methods measure the application circuit directly, but the introduction of errors and sweeping of the system clock means that they must be performed offline.

C. Arrival Time Detection

As the focus of reliability research has shifted from functional to timing faults, techniques have arisen that can detect late arrival at registers. The most well developed of these is Razor [7]. The scheme works by adding shadow registers to certain registers in a design, which clock the data input slightly later than the main register. If the shadow and main registers latch different values then a timing fault has occurred and action can be taken. Since these timing faults result in erroneous values being latched, mechanisms are required to correct these errors.

Similarly, it is possible to use shadow registers to detect impending timing faults: failure prediction. Here, the main and shadow registers share a clock, but additional delay is introduced into the shadow path. The amount of delay sets the guardband, the size of the window in which transitions are to be detected [8], [9]. If the data transitions within this guardband, a warning signal is generated that the device is operating close to the point of timing failure. An FPGA implementation of failure prediction is likely to use a secondary accelerated clock to drive the shadow registers, since finely-tunable delay elements are not as readily available as in ASICs.

D. Dynamic Voltage and Frequency Scaling

Device timing models are inherently conservative; they need to account for variations in process (including ageing), voltage and temperature (PVT). In practice, this results in circuits being operated at lower frequencies and at higher voltages than required. With suitable feedback, frequency and/or voltage can be adjusted at run-time to accommodate PVT variations and improve system capability or efficiency.

Basic implementations rely on pre-characterisation of the effect of PVT parameters [10] or inference from a TRC [11], [12], while more recent innovations make use of arrival time detection. The Razor architecture [7] is an example of dynamic voltage scaling implemented on CPUs. Timing faults are detected, and a closed-loop controller trades-off the error correction cost against supply voltage to achieve peak efficiency. The area overhead of circuitry needed to detect timing faults is relatively small since it is only applied to critical paths, but the associated error correction support must be applied to the entire clock domain and can be very costly.

Dynamic scaling additionally provides a means of circuit life extension. The predominant physical effects causing degradation, in current technologies, result in increased threshold voltage [13], which can be combated by DVS, and in slowed switching, which can be controlled by DFS.

III. Principle of Operation

The system presented herein improves on shadow register based arrival time detection methods in order to accurately measure timing slack, rather than provide just a pass/fail indication, at chosen nodes within a circuit. This is achieved by finely stepping, or sweeping, the phase of the shadow clock, detecting the exact point at which timing failure is induced in the shadow path. From this, the slack and maximum delay to the register can be calculated, facilitating in fine-grained monitoring of circuit health, and proportional control of DVS and DFS schemes.

Figure 1 shows the basic structure of the online measurement circuit, which can be applied to any register (P2) clocked by the main clock (M_CLK). The data input to the monitored register is forked off to a shadow register (S), which is clocked by a shadow clock S_CLK. S_CLK always operates at the same frequency as M_CLK, but its phase relationship (ϕ) is variable. The register outputs are compared to generate a mismatch signal (X), which is sampled by error register E. This signal is either directed to an error counter (which can be shared with other test sites using multiplexers) or stored as a first-fail flag. The paths being monitored (in this case between P1 and P2) are referred to as the Paths Under Monitoring (PUM). In practice, many paths end at the same register, and all of these are monitored by the addition of one shadow register.

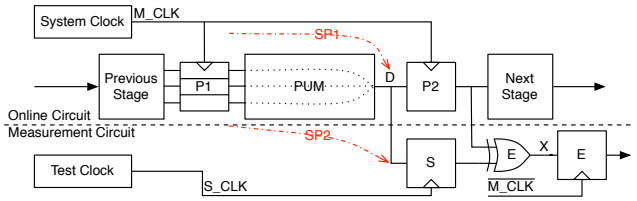


Figure 1: Basic principle of the online measurement circuit.

Figure 2 illustrates two possible cases: *1*) that the shadow path meets timing (passes), or *2*) does not (fails). In both cases, the slack of the combinatorial path is sufficient to meet the setup time for P2, so the circuit continues functioning correctly. In case 1, the shadow clock leads the main clock slightly, such that the rising edge occurs just before that of the main clock. There is still sufficient slack in path SP2 that the setup time of the shadow register is met. A glitch is visible at the XOR output since the two registers latch at different times, but this is not sampled by the error register. In case 2, the shadow clock leads further, such that path SP2 fails to meet the setup time of the S register. The disparity between the S and P2 registers is found by comparison, and the error is latched.

Connecting the error signal to a counter yields results as in Figure 3 which illustrates the count rate over a 360° sweep

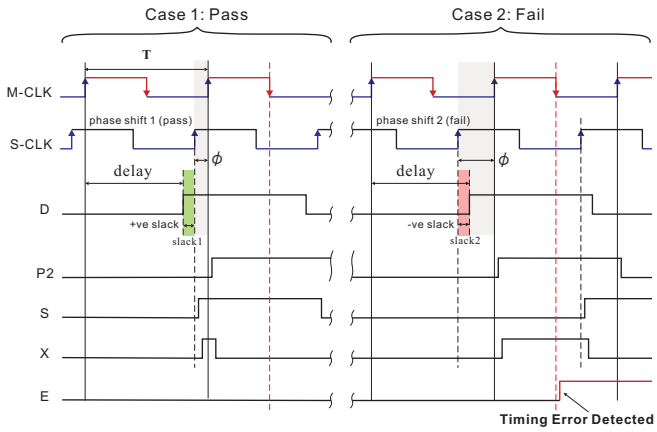


Figure 2: Timing diagram illustrating pass and failure of the shadow path.

of ϕ in a simple buffer chain. In region 1, E samples after M_CLK, ϕ is 0 or slightly positive and S and P2 will contain the same value, so no error is counted. As ϕ is increased further region 2 is entered. Here, register S will reach a point where its setup time is violated; it will sometimes latch a different value to P2 and errors occur. A 50% error count occurs halfway through region 2 due to the discrepancy between rising and falling delays in the PUM. Further still (region 3) and S will always sample one clock cycle behind P2, and an error will be recorded whenever P2 changes state.

The extent of region 1 gives the gross timing slack available: 1.90ns in this case, and by inference from the clock period, 6.67ns, the maximum delay from M_CLK to data arrival at S is (6.67ns – 1.90ns =) 4.77ns. The size and form of region 2 is determined by clock jitter and circuit complexity.

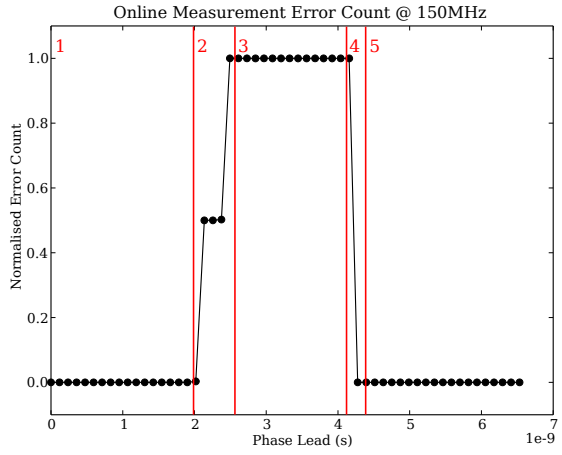


Figure 3: Example error profile for 12 LUT buffer cascade PUM, with regional annotations.

Table I: Location of blind spot for a different error clocks.

Clock for Register E	Blind Spot Location
M_CLK	Start of Sweep
S_CLK	End of Sweep
$\overline{\text{M\_CLK}}$	Mid-Right of Sweep
$\overline{\text{S\_CLK}}$	Mid-Left of Sweep

A. *Blind Spot*

The delay measurement scheme cannot measure over the entire range of the phase sweep; a small blind spot exists (region 4) due to the relative timing of P2, S and E registers. The choice of which clock and edge is used to trigger E determines the position of the blind spot. Table I illustrates the location of the blind spot for each of these; $\overline{\text{M\_CLK}}$ is used in Figure 3.

The blind spot represents the point where the advancement of ϕ causes S_CLK to wrap around and lag M_CLK rather than lead it. The error profile typically exhibits a discontinuity at this point.

The location of the blind spot is an important consideration, because if regions 2 and 4 overlap, timing measurement is impossible. Using the $\overline{\text{M\_CLK}}$ allows ϕ to exceed $+180°$ before reaching the blind spot. In addition, some capacity for measurement with lagging phase is maintained (region 5): essential for calibration, as discussed below.

B. *Calibration*

The technique described above assumes that SP1 and SP2 are identical and the clocks have no skew. In practice, this is not the case: the compilation flow is designed to ensure that optimisation priority is given to the application circuit, resulting in the shadow circuitry having a greater delay.

If the accuracy of measurement is not critical, the offset can simply be ignored; the insertion flow ensures that SP2 has a greater delay than SP1, and as such measurements will be conservative.

In order to improve accuracy, a novel calibration process has been developed that can measure the difference in path length between the main and shadow paths and is used to correct the online measurements, improving its accuracy. Path SP1 is measured offline using timing measurement techniques [4], [5], which requires no circuit modifications. S_CLK is configured to lag M_CLK such that S samples later then P2. The clock frequencies are gradually increased and at the point at which the path delay causes a violation of the setup time at P2, P2 will latch erroneous data, S will capture correct data and the error produced by comparison will be latched by E.

Once the maximum clock frequency for the PUM is found, the measurement offset $t_{SP2} - t_{SP1}$ is calculated, and this is used to correct the online measurements. Calibration can be carried out on power-up, or even at commissioning, and an offset stored for each test location.

C. Cost

FPGAs are an excellent fabric upon which to utilise this method of online measurement. Their sophisticated PLLs are able to generate the required clock signals, and extensive routing networks make it possible to route these and error signals with minimal overhead. When mapped to FPGA resources, two registers are required, one for the shadow and one the error. One LUT is needed for the comparison and an additional feeder LUT may be used in the shadow register path. In total two Logic Elements (LEs) are needed for each register monitored. The addition loading/fanout introduced by the shadow register has some delay implications, but the pre-connected nature of FPGAs means that the impact of this is minimal. There is also the cost of the error aggregation network, counters (if used) and control logic.

D. Reliability

The method must be robust to the factors that are under study: degradation, temperature and voltage. Circuitry after the main and shadow registers is not sensitive to small changes in delay; only correct data transmission is necessary. The path delay offsets and clock skew could be affected, however, the shadow and main paths are subjected to the same stimuli: experiments have shown that degradation is highly repeatable under given input conditions [13] and as such, any degradation to the shadow path will be very similar to that of the main path. Clock-trees are designed to minimise skew [5] and initial calibration compensates for any constant offset.

IV. FPGA Implementation

The technique described has been implemented on an Altera 65nm Cyclone III FPGA (EP3C16F484C6N) on a Terasic DE0 development board.

A. Clock Generation

PLLs are used to generate the main and shadow clocks. Since they have the same frequency, only one is required. The outputs have different phases: the main clock's is fixed, the shadow's programmable. The resolution of the measurement method is dictated by the size of the phase steps (a function of the PLL oscillator's frequency range), resulting in a best case of 96ps and worst 208ps in this device.

B. Results

In order to explore the basic principle, a simple application circuit, based on an adder, was monitored. The circuit input vectors are generated by an LFSR, and online measurement hardware has been added to the paths suggested to be most critical by the timing analysis (path 0 to 4). Both error counter and first-fail are used; also Transition Probability (TP) hardware is included to measure PUM delay independently and establish the accuracy of the technique.

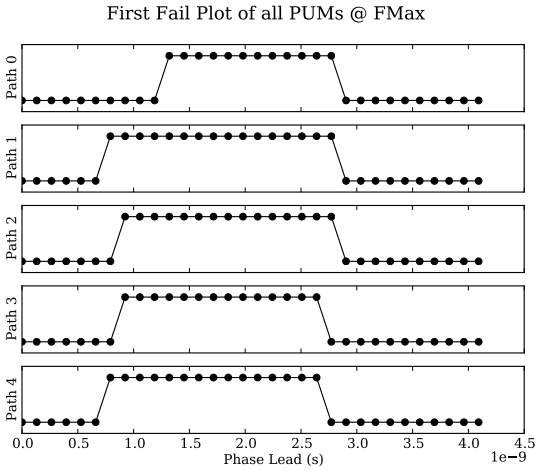


Figure 4: Plot of first-fail results for the five PUM.

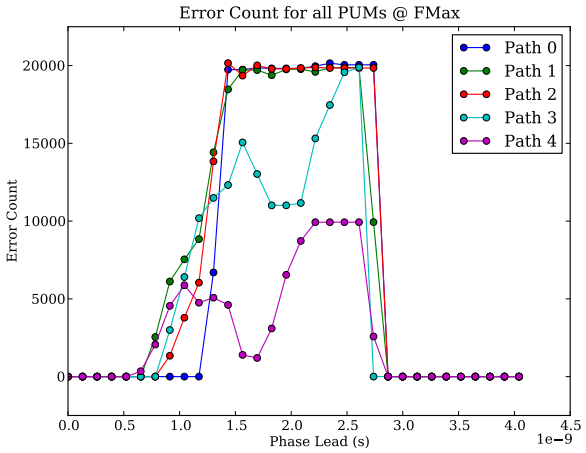


Figure 5: Plot showing the error profile for each of the five PUMs.

TP is a proven technique for establishing the delay of offline paths in arbitrary circuits [6].

Figure 4 shows results for the five PUMs in the example circuit using the simple first-fail circuitry. The different timing slacks at each register are visible: paths 1 and 4 appearing most critical, path 0 having the greatest slack. In Figure 5, the error count results are shown. This plot reveals the full complexity of the arrival time characteristics at the five registers.

Table II displays the CAD tool's timing model estimates for the path delays to the main and shadow registers. The shadow register delay is always longer as priority has been given to the main register to avoid impacting upon the application circuit's performance. Column 4 shows the TP measurements for the five main paths, taken under ambient temperatures. Note the large discrepancy between the timing

Table II: Timing Model Estimates for Main and Shadow Path Delays and Measured (TP) Main Path Delay.

PUM	Main (*ns*)	Shadow (*ns*)	TP Delay (*ns*)
Path 0	3.62	3.99	2.84
Path 1	4.12	4.64	3.24
Path 2	4.12	4.49	3.20
Path 3	4.24	4.61	3.29
Path 4	4.23	4.75	3.29

Table III: Delay Measured Using Transition Probability and Online Method Operating at F_{MAX}.

PUM	TP Delay (*ns*)	Online Delay (*ns*)	Error (*ps*)	Error
Path 0	2.84	2.96	124	4.4%
Path 1	3.24	3.59	356	11.0%
Path 2	3.20	3.38	186	5.8%
Path 3	3.29	3.49	204	6.2%
Path 4	3.29	3.59	347	10.7%

model delays and measured delays: on average the model is 30% slower. Importantly, while the absolute values of these are quite dissimilar, the relative order is the same for measured and modelled, meaning that the selection of registers to monitor can make use of the timing model.

The application circuit was run at the CAD tool's estimated F_{MAX} (236.47MHz) and the device's nominal core voltage. An uncalibrated online test was performed. At this frequency, the PLL configuration yields a phase step size of 105ps. These measurements are displayed in Table III, column 3. In columns 4 and 5, online and TP measurements are compared to establish the accuracy of the technique. The error is shown as the difference between the these delays, and as a percentage error. This error is made up of measurement error due to the lower resolution of the online method, and the shadow path length difference. Paths 1 and 4 have a greater error because there is greater delay in the routing to the shadow registers (shown in the timing model estimates of Table II).

The proposed calibration technique (III-B) was performed, and the average difference between the main and shadow path over a number of runs was subtracted from the measurement results. This drastically reduces the error, as shown in Table IV. The average error, when compared to the TP delay, is reduced to 0.58%. This represents the resolution induced measurement error, calibration removing the majority of path difference error.

C. Register Selection & Cost

The cost of implementing the additional circuitry for online measurement is related to the number of registers monitored. The choice and number of these registers in turn depends on the distribution of delays in the application circuit, and the coverage required. In order to explore the trade-off between area overhead and circuit coverage, an

Table IV: Calibrated Measurement Using Online Method Operating at F_{MAX}.

PUM	Delay (*ns*)	Error (*ps*)	Error
Path 0	2.80	-33.74	1.2%
Path 1	3.24	1.08	0.0%
Path 2	3.20	-2.45	0.1%
Path 3	3.32	34.82	1.1%
Path 4	3.23	-1.83	0.6%

experiment was conducted profiling the T20 benchmark set [14] (the twenty large MCNC benchmark circuits [15]).

The circuit netlists were wrapped in registers and compiled in Quartus II 11.1 for the Altera Cyclone III architecture using default compiler options. Timing analysis was performed to establish timing slack available at each of the internal and output registers, referred to as candidate registers (CREGs). Figure 6 illustrates a selection of slack distributions observed from the model.

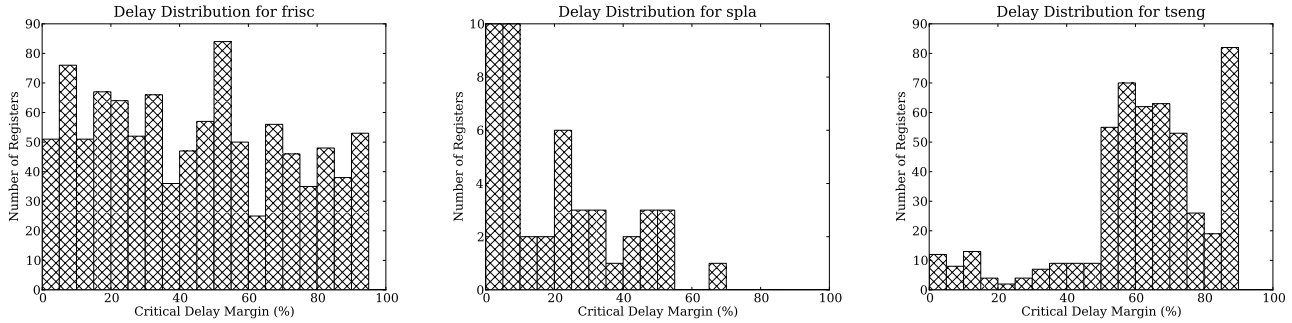
From the slack data given by the timing model it is possible to select registers to be monitored using a metric known as *Critical Delay Margin* (CDM). The timing model slack data can be used as it has been shown to be accurate for relative comparison of paths. For a particular CDM, CREGs are selected for monitoring if they fulfil the condition in Equation (1), where t_{prop} is the maximum propagation delay to the candidate register, t_{crit} is the maximum propagation delay in the entire circuit, and m the CDM, from 0% to 100%.

$$t_{prop} \geq t_{crit} \times (100 - m) \tag{1}$$

The degree of coverage required is a parameter to be chosen by the circuit designer; the decision dependent on possible variation in changes of delay between paths in the circuit. If it is assumed that all paths experience the same relative changes in delay, then only the single slowest path would need to be monitored, since it would always remain the critical path. However, given the variation observed in degradation rates alone [1], this approach would not be valid and timing failure would be likely to occur unnoticed.

Making the assumption that the worst case relative change in delay can affect any path independently, a coverage can be chosen to ensure that any path which would become critical when subjected to this delay increase is monitored. This is pessimistic since it makes the unlikely assumption that the fastest path monitored could slow down to beyond that of the critical path while all the slower paths monitored remain sub-critical, but provides an upper bound. It is expected that a CDM of 10% (approx. 11% maximum increase in delay) would be sufficient for sources of delay variation in current process technologies, however, further characterisation is required to confirm this selection.

Once the appropriate CDM for the circuit and technology is chosen, it is possible to determine the number of CREGs



(a) frisc (b) spla (c) tseng

Figure 6: Register slack distributions for three MCNC benchmarks, illustrating a range of timing distributions.

requiring monitoring and therefore the area overhead of the measurement circuitry. The delay distribution of the circuit is important: circuits with a distribution skewed to the left, such as Figure 6b, require the greatest number of CREGs to be monitored for low CDM as there are many near-critical paths. Circuits such as tseng (Figure 6c) have relatively few slower paths and so are adequately covered by monitoring just a few CREGs.

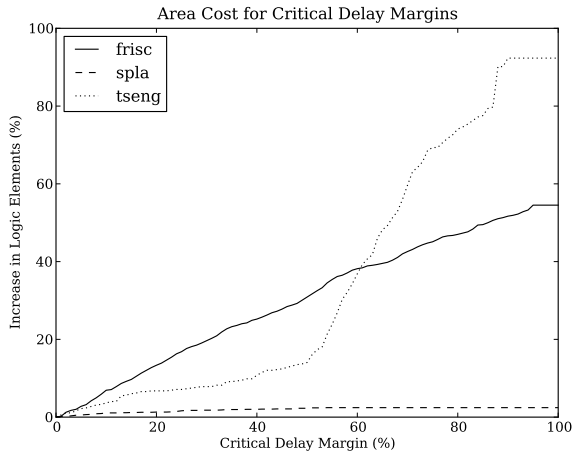


Figure 7: Area cost graph with CDM for the three delay distribution profiled benchmarks.

Table V shows the number of monitors, the number as a percentage of CREGs and relative cost as a percentage increase in LEs, for the twenty benchmark circuits investigated. The cost for a 10% CDM ranges from less than 0.08% to 8.3% with an average of 2.7%. The proportion of CREGs that are monitored depends only upon the delay distribution and choice of CDM. For example, spla, which has a distribution skewed towards slower paths, requires 43% of CREGs to be monitored, while tseng with a large number of fast paths needs only 3.9%. However, the LE overhead depends additionally on the LUT:REG ratio as register-rich designs inevitability require more monitors, and this factor dominates the difference between the benchmarks.

Figure 7 illustrates the relationship between LE cost and critical delay margin for the three benchmarks discussed earlier. Whilst the LE overhead of spla is small compared to the size of the circuit, most of this cost is incurred at low CDM. In frisc the overhead grows linearly reflecting the even distribution displayed in 6a, while the overhead in tseng increases sharply after 50% corresponding to the skew in its delay distribution.

To quantify the effect of the additional fanout/loading, due to the shadow register, on the main circuit paths, an experiment has been conducted using TP to measure the delay of a circuit with and without the shadow register. The delay increased from 3.51ns to 3.52ns, approximately 0.28%, a minimal increase is expected due to the highly connected nature of the FPGA fabric. The fanout/loading on particular LEs will vary slightly because of cluster depopulation, but this is believed to be representative of the order of the impact of shadow register addition on delay.

V. Dynamic Voltage/Frequency Scaling

A voltage/frequency scaling scheme making use of the demonstrated online measurement technique is presented in Figure 8. As with Razor, it is able to directly measure the result of changes in delay and so reduces the conservatism required by alternative methods. Unlike Razor, which controls the number of errors, the slack is measured and controlled in a closed loop to achieve a desired guardband. This removes the error correction overhead, which is large, particularly for the highly pipelined circuits frequently implemented on FPGAs.

A. Proof of Concept Implementation

The FPGA development board was modified, replacing the fixed core logic power supply with a variable one. In ambient conditions and at F_{MAX}, the register slack was monitored as the voltage was varied. This data is displayed in Figure 9. While maintaining a guardband slack of 145ps the

Table V: The cost, as a percentage increase in Logic Elements, for various Critical Delay Margins.

Benchmark	Base Resources			5% CDM		10% CDM		15% CDM		20% CDM	
	LE	CREG	LUT:REG	Monitors	**LE Cost**	Monitors	**LE Cost**	Monitors	**LE Cost**	Monitors	**LE Cost**
alu4	1522	8	69	3 (38%)	**0.39%**	4 (50%)	**0.53%**	6 (75%)	**0.79%**	6 (75%)	**0.79%**
apex2	1878	3	46	2 (67%)	**0.21%**	2 (67%)	**0.21%**	2 (67%)	**0.21%**	2 (67%)	**0.21%**
apex4	1261	18	47	5 (28%)	**0.79%**	13 (72%)	**2.1%**	17 (94%)	**2.7%**	17 (94%)	**2.7%**
bigkey	1923	421	2.6	7 (1.7%)	**0.73%**	54 (13%)	**5.6%**	139 (33%)	**14%**	217 (52%)	**23%**
clma	8416	101	52	2 (2.0%)	**0.048%**	3 (3.0%)	**0.071%**	5 (5.0%)	**0.12%**	6 (5.9%)	**0.14%**
des	1615	245	3.2	6 (2.4%)	**0.74%**	14 (5.7%)	**1.7%**	41 (17%)	**5.1%**	61 (25%)	**7.6%**
diffeq	1536	416	3.1	6 (1.4%)	**0.78%**	14 (3.4%)	**1.8%**	29 (7.0%)	**3.8%**	44 (11%)	**5.7%**
dsip	1591	421	2.1	9 (2.1%)	**1.1%**	66 (16%)	**8.3%**	150 (36%)	**19%**	210 (50%)	**26%**
elliptic	3718	1236	2.6	48 (3.9%)	**2.6%**	137 (11%)	**7.4%**	204 (17%)	**11%**	258 (21%)	**14%**
ex1010	4598	10	229	9 (90%)	**0.39%**	9 (90%)	**0.39%**	9 (90%)	**0.39%**	9 (90%)	**0.39%**
ex5p	1064	63	15	12 (19%)	**2.3%**	31 (49%)	**5.8%**	32 (51%)	**6.0%**	32 (51%)	**6.0%**
frisc	3672	1002	3.5	51 (5.1%)	**2.8%**	127 (13%)	**6.9%**	178 (18%)	**9.7%**	245 (24%)	**13%**
misex3	1397	14	50	6 (43%)	**0.86%**	13 (93%)	**1.9%**	13 (93%)	**1.9%**	13 (93%)	**1.9%**
pdc	4575	40	82	6 (15%)	**0.26%**	16 (40%)	**0.7%**	35 (88%)	**1.5%**	37 (92%)	**1.6%**
s298	1931	14	113	4 (29%)	**0.41%**	4 (29%)	**0.41%**	4 (29%)	**0.41%**	5 (36%)	**0.52%**
s38417	6467	1569	4	48 (3.1%)	**1.5%**	109 (6.9%)	**3.4%**	187 (12%)	**5.8%**	247 (16%)	**7.6%**
s38584.1	6586	1544	4.1	4 (0.26%)	**0.12%**	16 (1.0%)	**0.49%**	36 (2.3%)	**1.1%**	58 (3.8%)	**1.8%**
seq	1750	35	23	3 (8.6%)	**0.34%**	13 (37%)	**1.5%**	21 (60%)	**2.4%**	24 (69%)	**2.7%**
spla	3690	46	60	10 (22%)	**0.54%**	20 (43%)	**1.1%**	22 (48%)	**1.2%**	24 (52%)	**1.3%**
tseng	1096	507	1.9	12 (2.4%)	**2.2%**	20 (3.9%)	**3.6%**	33 (6.5%)	**6.0%**	37 (7.3%)	**6.8%**
Minimum					**0.048%**		**0.071%**		**0.12%**		**0.14%**
Maximum					**2.8%**		**8.3%**		**19%**		**26%**
Average					**0.95%**		**2.7%**		**4.7%**		**6.2%**

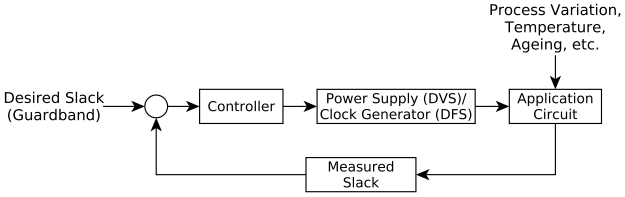


Figure 8: Proposed control loop for dynamic scaling.

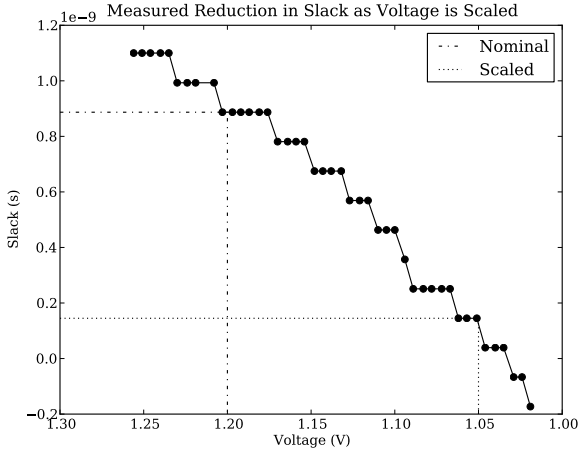


Figure 9: Relationship between core voltage and critical path slack.

circuit continued to operate down to 1.05V, reduced from the nominal core voltage of 1.20V. This demonstrates that circuit functionality can be maintained with estimated reductions of 23% in dynamic power, and 13% in static. The steps on the graph illustrate the resolution of the measurement method.

Since these experiments were performed at F_{MAX}, any power savings result exclusively from tuning for the particular device and environmental conditions. In practice, it is not uncommon for circuits to be operated at frequencies below F_{MAX}: here even greater savings can be made.

Alternatively, the excess slack can be reduced by DFS. The slack in the critical path at F_{MAX} is reduced from 889ps to a guardband of 145ps. This results in an operating frequency of 286.90MHz, an increase of 21%.

To establish the effect of temperature, one of the more dynamic influences on delay, timing measurements were taken as the package temperature was varied between 8°C and 130°C. Figure 10 illustrates that while delay varies by 291ps across this range, the rapid fluctuations are expected to be small and large changes gradual, so this should be easily controllable.

VI. Usage Considerations

A. Testing Architecture

It is possible to apply the online measurement hardware to application circuits in a number of ways. Each register monitored can have an error counter, or these can be shared among groups of monitors or all monitors. The counter can also be removed completely, using only first-fail latches. The phase can be swept continually, or it is possible to set a fixed phase guardband as in failure prediction and, when triggered, measure the slack at full resolution, to better track

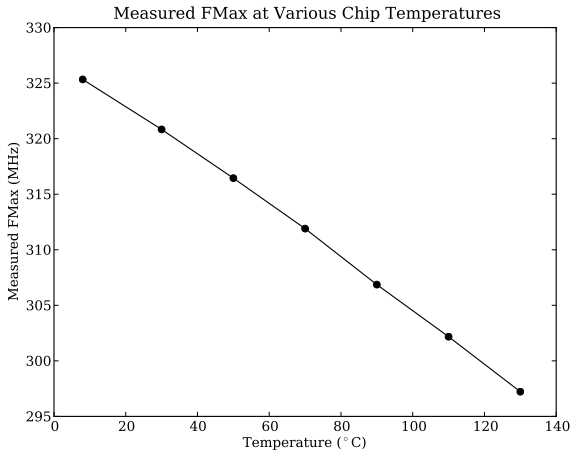


Figure 10: Relationship between maximum measured operating frequency and packaging temperature.

the progression to timing failure.

B. Unexercised Paths

Since measurement is conducted online, there is no control of the inputs into each PUM. It is possible that some paths will not be exercised in a test cycle, and as such the delay of these paths will not be measured. This is true of any non-exhaustive testing method and it affects accuracy. Consideration of this must be made when the measurement method is used for dynamic scaling. There is a possibility that the critical path may not be measured, and slack reduced. Doing so may cause the path to fail timing, introducing an error with no means of detecting or correcting it; small guardbands must be maintained to avoid this. Various methods, both experimental and statistical, are being investigated to overcome or reduce this problem.

VII. Conclusions

This work presents a technique for monitoring the performance of an online circuit through the measurement of timing slack, without disrupting the circuit's operation. This is demonstrated on an Altera Cyclone III FPGA, and compares favourably to measurements made using an existing, proven, offline method, with a maximum error of 1.2% when compared to that technique. A means of calibrating measurements is shown and a method for choosing which registers to monitor is proposed.

The performance impact of the measurement circuitry is shown to be small: just 0.28%. The cost of implementing additional hardware on a set of benchmarks is explored with an average increase in LEs of 2.7%, for sufficient coverage in current process technologies.

Initial experimentation of dynamic voltage and frequency scaling using the online measurement method have been investigated. Estimated power reductions of 23% dynamic and 13% static are achieved by DVS, or a performance increase of 21% using DFS.

Acknowledgements

The authors acknowledge the support of EPSRC (grants EP/C549481 and EP/H013784). They are also grateful for support from Altera.

References

[1] E. Stott, J. S. J. Wong, P. Sedcole, and P. Y. K. Cheung, "Degradation in FPGAs," in *ACM Intl. Symp. on Field programmable gate arrays*, 2010, pp. 229–238.

[2] M. Rothberg, "Disk drive for receiving setup data in a self monitoring analysis and reporting technology (SMART) command," U.S. Patent 6,895,500, 2005.

[3] J. Tschanz, K. Bowman, C. Wilkerson, S.-L. Lu, and T. Karnik, "Resilient circuits," in *ACM Intl. Conf. on Computer-Aided Design*, 2009, pp. 71–73.

[4] J. S. J. Wong, P. Sedcole, and P. Y. K. Cheung, "Self-Measurement of Combinatorial Circuit Delays in FPGAs," *ACM Transactions on Reconfigurable Technology and Systems*, pp. 10:1–10:22, 2009.

[5] P. Sedcole, J. S. J. Wong, and P. Y. K. Cheung, "Characterisation of FPGA Clock Variability," in *IEEE Comput. Soc. Symp. on VLSI*, 2008, pp. 322–328.

[6] J. S. J. Wong, P. Sedcole, and P. Cheung, "A transition probability based delay measurement method for arbitrary circuits on FPGAs," in *Intl. Conf. on Field-Programmable Technology*, 2008, pp. 105–112.

[7] D. Ernst *et al.*, "Razor: a low-power pipeline based on circuit-level timing speculation," in *IEEE/ACM Intl. Symp. on Microarchitecture*, 2003, pp. 7–18.

[8] M. Agarwal, B. C. Paul, M. Zhang, and S. Mitra, "Circuit Failure Prediction and Its Application to Transistor Aging," in *IEEE VLSI Test Symp.*, 2007, pp. 277–286.

[9] A. Amouri and M. Tahoori, "A Low-Cost Sensor for Aging and Late Transitions Detection In Modern FPGAs," in *IEEE Intl. Conf. on Field Programmable Logic and Applications*, 2011, pp. 329–335.

[10] T. Burd, T. Pering, A. Stratakos, and R. Brodersen, "A dynamic voltage scaled microprocessor system," *IEEE Journal of Solid-State Circuits*, pp. 1571–1580, 2000.

[11] C. Chow, L. Tsui, P. Leong, W. Luk, and S. Wilton, "Dynamic voltage scaling for commercial fpgas," in *IEEE Intl. Conf. on Field-Programmable Technology*, 2005, pp. 173–180.

[12] M. Nakai *et al.*, "Dynamic voltage and frequency management for a low-power embedded microprocessor," *IEEE Journal of Solid-State Circuits*, pp. 28–35, 2005.

[13] E. Stott, J. S. J. Wong, and P. Cheung, "Degradation analysis and mitigation in fpgas," in *IEEE Intl. Conf. on Field Programmable Logic and Applications*, 2010, pp. 428–433.

[14] V. Betz and J. Rose, "Vpr: A new packing, placement and routing tool for fpga research," in *Intl. Workshop on Field-Programmable Logic and Applications*, 1997, pp. 213–222.

[15] S. Yang, "Logic synthesis and optimization benchmarks user guide version 3.0," 1991.