VaROT: Methodology for Variation-Tolerant DSP Hardware Design Using Post-Silicon Truncation of Operand Width

Keerthi Kunaparaju, Seetharam Narasimhan, Swarup Bhunia

Dept of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, USA
{kxk239, sxn124, skb21}@case.edu

Abstract— **With increasing parameter variations in nanoscale technologies, computational blocks in Digital Signal Processing (DSP) hardware become increasingly vulnerable to variation-induced delay failures. These failures can significantly affect the Quality of Service (QoS) for a DSP chip leading to degradation in parametric yield. Existing post-silicon calibration and repair approaches, which rely on adaptation of circuit operating parameters such as voltage, frequency or body bias, typically incur large delay or power overhead in order to maintain QoS. In this paper, we present a novel low overhead approach of healing DSP chips by commensurately truncating the operand width based on its process corner. The proposed approach exploits the fact that critical timing paths in DSP datapaths typically originate from the least significant bits (LSBs). This condition can also be satisfied by skewing the path delay distribution during logic synthesis or gate sizing. Hence, truncation of the LSBs, realized by setting them at constant values, can effectively reduce the delay of a unit, thereby avoiding delay failures. We also note that truncation of LSBs typically has minimal impact on QoS. Besides, efficient choice of truncation bits and values can minimize the impact on QoS. We propose appropriate design time modifications including insertion of low-overhead truncation circuit and gate sizing to maximize the delay improvement with truncation. Simulation results for a Discrete Cosine Transform (DCT) application at 45nm technology show large improvement in yield (41.6%) with up to 5X savings in power compared to existing healing approaches.**

Keywords-Variation Tolerance, DSP, Operand Truncation, Low Power Operation

I. INTRODUCTION

Increasing process parameter variations cause large spread in major circuit parameters such as speed and power consumption which significantly affects the manufacturing yield [1, 9]. Since worst-case design and statistical design [2, 3] approach may incur a large design overhead in terms of area, power and performance, designers resort to two major design techniques to ensure high yield under parameter variations at low design overhead: 1) *Variation Tolerant Design Approaches [6]*, where circuits are designed to account for process variations during run time such that the performance of the chips will not be affected. But this might lead to increase in design complexity and also incur more area and power overhead; 2) *Post-silicon calibration and repair*, where parameter shift is detected and compensated after manufacturing by changing operating parameters such as supply voltage, frequency or body bias [4, 5, 8]. But scaling up the supply voltage results in high power consumption due to quadratic dependence of dynamic power of a circuit on operating voltage. Also application of body bias increases leakage current.

In this paper, we propose *VaROT* – a **Va**riation **R**esilience through **O**perand **T**runcation approach targeting yield improvement in digital signal processing (DSP) hardware. VaROT provides a novel, low-overhead approach for post-silicon healing of Integrated Circuits (ICs) which suffer from delay failures to restore system performance under large die-to-die or within-die parameter variations. The proposed approach exploits the fact that in typical DSP datapath modules (such as adder, multiplier, multiply-and-accumulate units), critical timing paths originate from the least significant bits (LSBs) and they can be shortened by truncating the LSBs – i.e. setting constant values e.g. "0" to these bits. Consequently, truncation of operand width in these data paths post-manufacturing can be used to avoid delay failures in most significant bits (MSBs). Truncating the input bits, however, affects the output quality but it is found that in common DSP computations (such as filtering, Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), color interpolation, motion estimation), truncating the LSBs lead to minimum loss in output quality of service (QoS) [6, 12, 13]. Note that given a DSP datapath, the effect of truncation on delay reduction can be maximized by using a constrained design optimization step, e.g. gate sizing, which ensures that the critical paths originate from the input bits which have minimum impact on QoS. Besides, one can choose the optimal combination of constant values that can be assigned to the truncated bits to minimize QoS impact. During design of a DSP system, the truncation hardware can be inserted in the datapath circuits. Post-silicon, the truncation circuit is turned on and appropriate numbers of input bits are truncated to the predetermined truncation values, based on the sensed process corner of a chip. Unlike the existing post-silicon repair solutions e.g. voltage or frequency scaling, simulation results show that such healing procedure avoids large impact on power dissipation, area overhead and throughput improving yield.

The novelty of this technique lies in applying *dynamic* bit width truncation post manufacturing because by applying

bit width truncation during design time, the ICs meeting the clock constraint will also undergo truncation which is not absolutely necessary resulting in quality degradation in all ICs. We refer to it as dynamic truncation since the number of bits truncated depends on the process corner of the IC. Unlike the finite word length approach in [7], the paths in the design are skewed such that critical paths originate from LSBs. In particular, the paper makes the following contributions:

1. It presents a methodology for increasing the efficiency of dynamic truncation in variation-resilient DSP circuits.
2. It also presents a low-overhead implementation of the truncation hardware.
3. Considering a common DSP application, namely Discrete Cosine Transform (DCT), it verifies the effectiveness of the proposed approach in improving parametric yield without high power overhead.

II. MOTIVATION

Increase in threshold voltage due to process variations increases the delay inside the circuit [9] which can lead to delay failures, where all the output bits are not computed correctly within the clock constraint and such ICs are usually discarded post-manufacturing. However, we note that certain DSP applications can tolerate some error in their outputs as long as the error is within an acceptable margin, determined by the Quality of Service (QoS) of the application. This is possible if the failing bits are not the most significant bits (MSBs) of the outputs and they can be computed correctly within the clock constraint if they do not fall on the critical path. This motivated us to investigate techniques for healing the ICs without affecting the power by on-demand removal of the critical path instead of reducing its delay. In order to "remove" the critical path or prevent it from being excited, we truncate its input so that the critical delay shifts to that of the next critical path. However this delay reduction could be at the cost of incorrect computations at some of the output bits but much better than the process variation induced delay failures, which typically affect the MSBs. Effect of truncation on delay and quality is discussed below through the simulation results of 2 bit adder and 8 bit adder respectively.
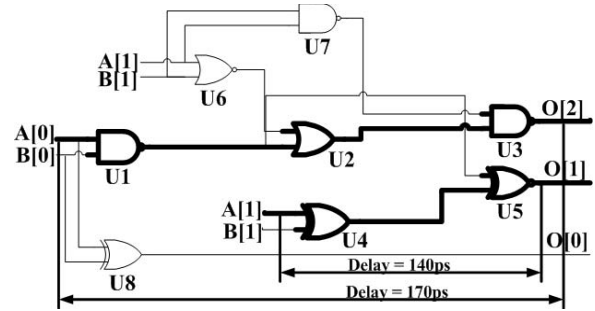

Figure 1. Effect of truncation on critical path delay for 2 bit adder.

A. Effect of Truncation on Critical Path

We synthesized a 2 bit adder using *Synopsys Design Compiler* and IBM 90nm standard cell library. The gate-level circuit is shown in Fig. 1. It is observed that by truncating A[0] the critical path which is from A[0]->O[2] with a delay of 170ns is shifting to the next highest path starting from A[1]->O[1] with a reduced delay of 140ps.

B. Effect of Truncation on Quality

The impact of truncation on quality is seen by simulating a two input 8 bit adder. The outputs of the adder under 3 different cases i.e. Case 1 without process variations, Case 2 with 10% inter-die variations and Case 3 with both variations and truncation applied are shown in Table 1. In Case 2, due to process variations the MSBs A[7] and B[7] failed to latch the correct data within the target delay whereas in Case 3, truncation of LSBs A[0] and B[0] prevented the excitation of the critical path and all the output bits are computed correctly within the clock constraint as shown in the table. The output decimal value in this case with truncation is 255 which is much better than 128 in Case 2 with process variations and is very close to original value 256 in Case 1. Thus truncation has a minimum impact on the output quality allowing MSBs to compute correctly within the clock constraint. The simulation results for 8 bit adder and multiplier are presented in Tables 2 and 3 respectively and it is observed that as we truncate more input LSBs, we get higher delay reduction. In other words, to accommodate for the higher increase in delay due to process variations we can truncate appropriate number of input bits.

Table 1. Effect of variation with and without truncation on 8-bit adder

Critical path	Without Process Variations (Case 1)		With Process Variations (Case 2)		With Process Variations and Truncation (Case 3)	
	Delay (ps)	Addition Computation	Delay (ps)	Addition Computation	Delay (ps)	Addition Computation

$B(0) \rightarrow O(8)$	145	Carry 1111111
		A 10101011
		B 01010101
		        100000000
			154	Carry 1111111
				A 10101011
				B 01010101
				      10000000
					Truncated A[0] B[0].Critical path not excited and delay is less than 145ps	Carry 0000000
						A 10101010
						B 01010100
						      11111110

Output Decimal Value		256		128 (at 145ps)		255

Table 2. Truncation Results for 8 bit adder		Table 3. Truncation Results for 8 bit multiplier	
# of bits truncated	%decrease in delay	# of bits truncated	%decrease in delay
2	12.3	2	2.6
3	25.8	4	9.6
4	39.3	6	30.8
5	51.6	8	37.1

III. METHODOLOGY

The truncation-based healing methodology can be applied to heal any DSP circuit where we can trade-off QoS to increase manufacturing yield with minimal power overhead. The two main features of this technique are:

- Truncation of least significant input bits has less impact on QoS and allows the circuit to meet the delay target.
- Truncation also helps in saving some switching power as it eliminates switching activity at the truncated nodes.

The proposed methodology is shown in Fig. 2 using a flow chart. It is primarily classified into two phases: 1) Design Phase and 2) Manufacturing Test Phase. For a given design, the inputs are the target delay constraint (D_{max}) and a set S of different frequency bins into which the manufactured ICs can be classified, post-manufacturing. The output is healed chips meeting the target delay which are sorted into different bins based on their QoS. Note that VaROT can be used as alternative or complementary to existing design-time approaches [6] where the most significant coefficients are computed with higher delay margins. Next we describe each of the steps in detail.

A. Design Phase

1) Perform Timing analysis and Sizing: Timing analysis is performed to find the delays of the paths originating from all the primary input bits and if the longest paths in the design do not originate from LSBs, appropriate sizing constraints are applied. Tighter delay constraints are set on the paths originating from MSBs and relaxed delay constraints are set on the paths originating from LSBs. The paths should be skewed such that when the highest delay path is truncated we get large delay reduction. For each frequency bin in the input set S, the amount of delay tolerated by each bin and the paths to be truncated to compensate for that delay are found.

2) Apply Truncation: Truncation values are assigned to the input bits to truncate the paths and the impact on the output quality is seen by simulating the netlist. For example, for a particular frequency bin to tolerate the delay if two input bits have to be truncated to shift the critical path to the next highest path then combination of all truncation values for those input bits i.e. 00, 01, 10, 11 are applied and the optimal combination which has a minimal impact on the quality, while meeting the required delay tolerance, is selected. Thus, the optimal truncation values are determined

for each frequency bin and if the impact of truncation values on QoS exceeds an acceptable QoS margin, then we stop applying truncation.

3) Truncation circuit: It is designed with a minimum overhead and needs to have a provision to truncate different number of input bits to heal different ICs in different frequency bins. To prevent the critical path excitation we can either truncate the *inputs* or the *outputs* the first level gates. One obvious way to perform truncation of the input bits is to insert NAND/NOR gates which can gate those bits to '0' or '1'. The other way is setting/resetting the individual input bit flip-flops and clock gating them. However, both schemes incur huge area and delay overhead. So we decided to truncate the outputs of the gates. To avoid the leakage using only single pull-up or pull-down transistor, the first-level gates can be supply/ground-gated when the pull-down/pull-up transistors are turned on for truncating the gate outputs [11]. If a particular input bit going to any inverting gate has to be truncated to constant '1', VDD gating is applied at the output of that gate and a pull down
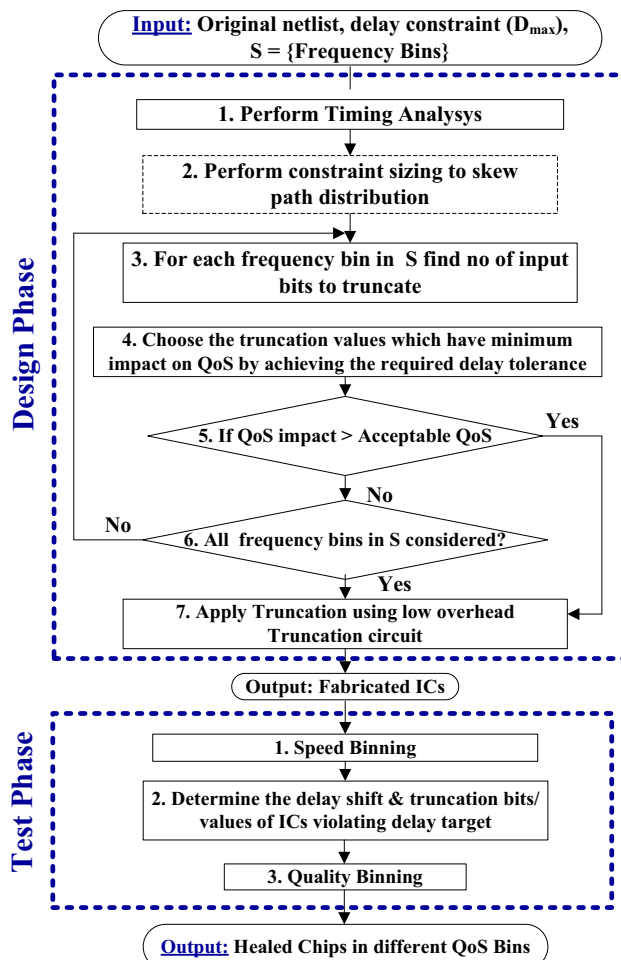

Figure 2. Flow chart of the design and test methodology for the proposed truncation approach.
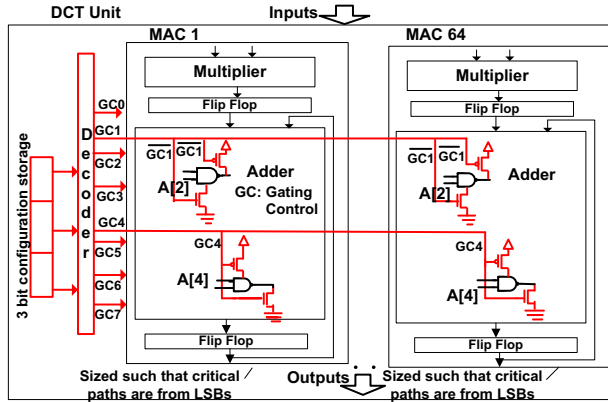
Figure 3. DCT Hardware with truncation scheme.

transistor is used to force the output to GND as shown in Fig. 3. Inside a multiply-and-accumulate (MAC) unit of the DCT architecture, truncation of input bit A[2] to constant '0' is performed by applying ground-gating and pull-up transistor at the gate output. Similarly for an input bit whose value has to be truncated to constant '1', supply-gating is applied and the output of that gate is forced to GND using a pull-down transistor as shown in Fig. 3 for the A[4] bit. The gating, pull-up and pull-down transistors are controlled by the *gating control* (GC) signals whose value will be high only in the truncation mode. The gating control signals for achieving different levels of truncation are generated using a decoder which is driven by some configuration bits stored in a non-volatile memory. These bits can be fixed during manufacturing test phase for different ICs which will be truncated at different levels. One of the input combinations of the decoder corresponds to no truncation, which will be applied to the chips which already meet the target delay, post-manufacturing. This scheme ensures minimal area overhead, caused by decoder circuit and 2 extra transistors for each first-level gate which needs to be truncated.

B. Manufacturing Test Phase

The truncation circuits have to be incorporated in all the ICs during design phase. Post-manufacturing, the ICs are subject to testing and speed binning to sort the ICs in different bins based on their frequency of operation. To compensate for the delay increment, appropriate input combination of bits should be stored in an on-chip/on-board non volatile memory so that appropriate truncation is always applied to the IC making it always meet the clock constraint. However, it should be noted that the healed ICs now fall into nominal frequency bins but provide different QoS levels depending on the truncated bits. In the final step, quality binning is performed and the healed ICs are distributed in different bins based on the amount of quality degradation. Apart from compensating for process variations, other use of the proposed healing approach is to compensate for aging-induced temporal variations where the ICs need to be periodically tested and characterized and are

healed by applying appropriate truncation as long as the quality degradation can be tolerated.

IV. RESULTS

We applied VaROT to a DSP circuit, commonly used in image processing and video compression, namely Discrete Cosine Transform (DCT) [12]. The design of a 2-D Discrete Cosine Transform was obtained from [14]. It takes as its input an 8x8 block of 10-bit pixels from an image and outputs sixty-four 12-bit DCT coefficients. The DCT architecture used has 64 MAC units which compute each DCT coefficient in parallel. A MAC unit consists of a 24-bit multiplier followed by a 27-bit adder in different pipeline stages as shown in Fig. 3. As all MAC units run in parallel the critical path is through a single MAC unit and as it is pipelined the critical path is through the larger adder block. The DCT design is synthesized with a clock constraint of 3.5ns using *Synopsys Design Compiler* and mapped to IBM 90nm standard cell library. The target is to improve the yield by healing bins with input set S = {3.6ns, 3.71ns, 3.85ns, 4.0ns, 4.16ns}. By following the design flow as shown in Fig. 2, static timing analysis is performed on the gate-level netlist and sizing constraints are applied to the DCT design such that critical paths originate from LSBs of the adder since truncating the input bits of the multiplier is resulting in more quality loss. For each frequency bin in S, the amount of delay increment is calculated. In this example, the first bin exceeds the nominal delay by 3% and it is observed from timing analysis that to compensate, three input bits (A[0], A[1], B[1] assuming A and B as inputs of the adder) have to be truncated so that the critical path shifts to the next highest path (originating from A[2]) and the optimal truncation combination '000' is applied. Finally, the selected truncation values to the input bits are implemented using truncation circuit. A 3-to-8 decoder is used to apply different levels of truncation from 3 to 9 bits and one of the input combinations is designed to cause no truncation. The truncation bits and their corresponding values are

Figure 4. a) Original Image; b) Output image with variations; c) Output image with variations and truncation.

Table 4. Truncation Results for DCT Design

# of truncated bits	% of delay reduction	PSNR (db)							% reduction in power
		Lena (512x512)	Kiel (512x512)	Barbara (512x512)	Lake (512x512)	Clown (256x256)	Aero (256x256)	House (256x256)	
Original	------	50.83	52.31	51.16	48.71	48.84	50.78	51.08.	--------
3	3.50	50.81	52.27	51.13	48.69	48.82	50.71	50.99	0.28
4	6.70	50.77	52.20	51.08	48.67	48.80	50.69	50.94	1.40
5	10.06	50.69	52.01	50.94	48.59	48.71	50.70	50.80	1.92
6	13.41	50.5	51.72	50.63	48.41	48.54	50.42	50.47	2.66
7	16.46	50.04	51.15	49.97	48.01	48.09	49.89	49.81	3.33
8	19.81	48.93	49.68	48.54	46.99	47.07	48.72	48.28	4.03
9	22.86	46.45	46.66	45.61	44.73	44.80	46.04	45.37	4.76
10	**26.21**	**42.37**	**42.34**	**41.26**	**41.87**	**41.81**	**43.74**	**42.89**	**4.90**

determined for all frequency bins in set S. Table 4 lists the percentage decrease in delay for different number of input bits truncation, the impact on quality for each truncation measured in terms of PSNR on different images like Lena, Kiel, Barbara, Lake, Clown, Aero and House and percentage decrease in switching power for each truncation as switching activity at some nodes is reduced. Thus, first two columns of Table 4 serve as a reference for the designer, after the chips are manufactured, to see how many input bits have to be truncated to compensate for a particular delay increment in order to heal the failing ICs. By estimating the delay and area, comparison is made between original architecture and the architecture with truncation circuit as shown in Table 5. The values show that critical path delay of the architecture with truncation circuit has only 1.2% overhead. The area overhead due to pull-up, pull-down and gating transistors is less than 1% and since they will only switch once; hence there will be no dynamic power overhead. In fact the switching power decreases due to decrease in input switching activity as more input bits are truncated, as shown in Table 4. However, for a chip with no truncation applied, the power overhead is due to the extra leakage caused by the decoder and truncation transistors.

The effect of process variations on the output image quality before and after applying truncation is observed by simulating the DCT design and the quality impact on Lena image is shown in Fig. 4. From Case 1 in Fig. 4 it is observed that quality of the image is much better and even close to the original image after healing by means of truncating bits. But in Case 2, to compensate delay for extreme process variations the number of bits to truncate increases and the output quality degrades significantly.

A. Impact on Manufacturing Yield

By using the proposed approach yield can be increased significantly with minimal area and power overhead at the cost of slight degradation in output QoS. We performed Monte Carlo simulations for the DCT circuit in HSPICE using PTM45nm technology [15] for 10,000 process corners with inter-die variation of 20% and intra-die variation of 15%. The

Table 5. Comparison of area and delay

	Original	VaROT	Overhead	
			VaROT	Existing Method [6]
Delay (ps)	407	412	1.20%	11.56%
Area (mm^2)	9.92	10.02	0.96%	12.23%

resulting delay distribution histogram is shown in Fig. 5. By defining the QoS margin of the healed ICs to be less than 3dB compared to nominal ICs, truncation till 9 bits is performed and it is found that yield significantly improved from 51.6% without truncation to 93.2% after truncation. The corresponding quality bins are also shown in Fig. 5. Thus the healed ICs fall in the bins with different degrees of quality loss and depending on the customer requirement of acceptable QoS margin, the healed ICs can be salvaged.

B. Power Savings with VaROT

Next we compare the power savings achieved with our technique when compared to supply voltage scaling and body biasing-based healing techniques. It is known that process variations increase the threshold voltage of transistors and so ICs which need healing consume low power compared to the nominal ICs. This fact is exploited by the truncation technique to achieve low power healed ICs compared to nominal ICs whereas other techniques require
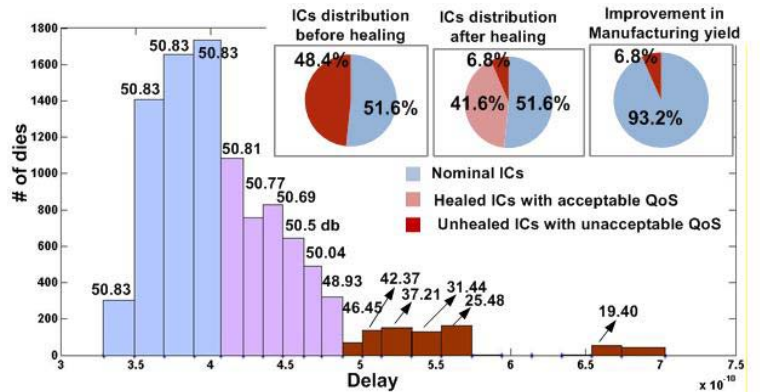

Figure 5. Post-manufacturing delay distribution of 10,000 dies. By using truncation, chips in different frequency bins can be healed leading to increased yield. However, these healed ICs fall into degraded but acceptable QoS bins. The chips which cannot be healed within acceptable QoS margin still lead to yield loss of 7%.

Table 6. Power savings with Truncation compared to two alternative approaches: VDD scaling and FBB

% Delay shift	Vdd Scaling (V)	Vdd Scaling Power (mW)	Optimal Body Bias (FBB)(V)		FBB Power (mW)	VaROT # of bits Truncated	% loss in QoS	VaROT Power (mW)	Power Savings(%)	
			Vb PMOS	Vb NMOS					Over Vdd Scaling	Over FBB
0	1.0	19.22	1.0	0.0	19.22	0	0	19.22	0	0
+3.22	1.11	26.55	0.6	0.2	20.90	3	0	17.53	51	19
+6.52	1.13	26.21	0.5	0.5	21.83	4	0.001	16.06	63	35
+9.95	1.19	28.39	0.3	0.85	22.39	5	0.003	14.68	93	52
+12.60	1.24	31.61	0.25	0.9	23.45	6	0.01	13.69	131	71
+15.57	1.29	34.46	0.22	0.93	27.25	7	0.02	12.73	171	114
+19.67	1.38	41.58	0.20	0.95	33.26	8	0.04	11.52	261	188
+22.82	1.44	47.65	0.19	0.98	37.96	9	0.1	10.66	347	256
+23.95	1.46	50.00	0.18	1.0	46.10	10	0.2	10.38	382	344

extra power consumption for healing. We calculated the power savings by simulating the DCT design in HSPICE and applied voltage scaling and body biasing techniques. In case of body biasing, we used Forward Body Biasing (FBB) since it is the most effective way to reduce both active and leakage power and improve the performance of the circuit. Table 6 lists the percentage increment in power consumption (compared to the nominal power) due to healing the ICs at different process corners by scaling up the V_{DD} and body biasing to meet the target delay. The table also lists the percentage increase in power savings that can be achieved with our technique for the same improvement in yield over voltage scaling and FBB, the number of bits to be truncated to compensate for the delay and loss in QoS at every truncation level. The table shows large power savings can be achieved through VaROT when compared to voltage scaling and FBB techniques at the cost of slight loss in QoS with significant increase in manufacturing yield. Though there is a little impact on the output quality, the designer depending on the demand for output quality can always limit the number of truncation bits.

V. CONCLUSION

We have presented *VaROT* - a low-overhead post-silicon compensation approach for DSP hardware using dynamic truncation of operand width. The proposed approach can improve the profit with minimal impact on QoS. It exploits the fact that critical paths in DSP datapaths typically originate from the input LSBs and truncation of these bits to fixed values results in shortening of the timing paths, leading to avoidance of delay failures in slow process corners without affecting the QoS considerably. The approach can be effective for generic DSP applications which exhibit these properties. Unlike the existing healing approaches using voltage/frequency scaling or body biasing, the proposed approach does not affect the performance and power of the DSP chips. The paper presents a design methodology to minimize the overhead due to truncation hardware and a gate sizing step to maximize delay improvement with truncation. Simulation results for an example DCT application demonstrate the effectiveness of the approach in improving parametric yield by repairing the chips which fail to meet the target delay due to variations. The healed ICs however suffer from slight degradation in QoS over nominal value. The proposed approach, hence, can benefit from a quality binning step, which sorts the repaired chips in bins with acceptable but slightly degraded QoS. Although we use truncation for process compensation, it can also be effective for dynamic adaptation to temporal variations – e.g. aging or environment induced variations.

ACKNOWLEDGMENT

REFERENCES

[1] K. A. Bowman *et al*, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration", *IEEE JSSC*, 2002.
[2] H. Chang and S.S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single pert-like traversal", *ICCAD*, 2003
[3] A. Agarwal *et al*, "Circuit optimization using statistical timing analysis", *DAC*, 2005.
[4] J.W. Tschanz *et al*, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage", *IEEE JSSC*, 2002.
[5] J.W. Tschanz *et al*, "Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors", *IEEE JSSC*, 2003.
[6] N. Banerjee, G. Karakonstantis and K. Roy, "Process variation tolerant low power DCT architecture", *DATE*, 2008.
[7] Y. Liu *et al*, "Design of low-power variation tolerant signal processing systems with adaptive finite word-length configuration", *ISQED*, 2010.
[8] S. Narendra *et al*, "Impact of using adaptive body bias to compensate die to die Vt variation on within-die Vt variation", *ISLPED*, 1999.
[9] S. Borkar *et al*, "Parameter variations and impact on circuits and microarchitecture", *DAC*, 2003.
[10] A. Datta *et al*, "Speed binning aware design methodology to improve profit under parameter variations", *ASP-DAC*, 2006.
[11] S. Bhunia *et al,* "Low-power scan design using first-level supply gating", *IEEE TVLSI*, 2005.
[12] R.C. Gonzalez *et al*, *Digital Image Processing*, Prentice Hall, 2002.
[13] R. Hedge and N.R. Shanbhag, "Soft digital signal processing", *IEEE Trans. VLSI*, vol. 9, no. 6, pp. 813-823, 2001.
[14] Open Cores [Online] http://www.opencores.org
[15] Predictive Technology Model [Online] http://www.eas.asu.edu/~ptm