DVAS: Dynamic Voltage Accuracy Scaling for Increased Energy-Efficiency in Approximate Computing

Bert Moons, Marian Verhelst
Department of Electrical Engineering - ESAT, KU Leuven, Leuven, Belgium
Contact: bert.moons@esat.kuleuven.be

***Abstract*—A wide variety of existing and emerging applications in recognition, mining and synthesis and machine-to-human interactions tolerate small errors or deviations in their computational results. Digital systems can exploit this error tolerance to increase their energy efficiency, which is crucial in high performance wearable electronics and in emerging low power systems for the internet-of-things. A dynamic energy-accuracy trade-off brings an extra degree of freedom for system level power management. We introduce the concept of Dynamic Voltage Accuracy Scaling and illustrate its analogy to Dynamic Voltage Frequency Scaling. Dynamic Voltage Accuracy Scaling proves to have higher energy gains at most output qualities compared to other approximate computing alternatives. This work further generalizes the Dynamic Voltage Accuracy Scaling concept to pipelined structures and quantifies its energy overhead. Shallow pipelined multipliers with two to four dynamic accuracy modes can be supported with limited ($< 10-20\%$) overhead, resulting in significant energy savings of up to 90% or more for less than 2% mean error. DVAS is finally applied to a JPEG image processing application, demonstrating large system level gains without noticeable impact to user or application.**

***Index Terms*—approximate computing, voltage scaling, DVAS, Dynamic Voltage Accuracy Scaling, scalable effort, approximate multiplier**

I. Introduction

A wide variety of applications such as recognition, mining and synthesis (RMS), or human-to-machine (H2M) and machine-to-human (M2H) interactions tolerate small errors or deviations in output quality and are inherently fault-tolerant. This observation has led to a new class of hardware design techniques known as *approximate computing* [1] [2]. These techniques let digital processing approximate results, allowing hardware to trade-off energy for accuracy. A processor could as such invest a lot of energy only when very accurate data processing is required and save energy when less accurate processing is permitted. This effectively reduces average energy consumption and creates an extra degree of freedom for system-level power management. The earliest work allows this trade-off to be installed in digital arithmetic at design time [3] [4] [5], while more recent work is focused on at run-time adaptable systems [6] [7] [8]. Note that this idea of dynamic scaling is analogous to the operation of the human brain, which can selectively increase its efforts when more difficult or precise computations are needed.

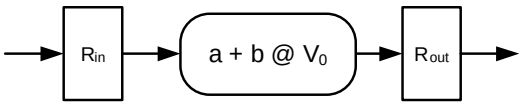
The targeted applications are omnipresent in several wearable systems, which are powered by batteries (e.g. smartphones or Google's project Glass) or energy-harvesting (e.g. internet-of-things sensor nodes). Both options imply operation with a very scarce power-budget. Making energy-efficiency crucial for system lifetime and user satisfaction [9].

The energy-efficiency of e.g. digital multipliers can be increased by truncating or rounding its inputs to reduce the circuits switching activity [7] [10]. This introduces small deviations in output quality caused by higher quantization errors on the inputs and thus decreases accuracy. However, using smaller bit widths not only reduces circuit activity, but also shortens data paths and thus allows voltage scaling without timing induced errors. This is in contrast to [6] and [11]. We introduce *Dynamic Voltage Accuracy Scaling* (DVAS) as a term comprehending all dynamic techniques that allow voltage scaling through accuracy reduction without introducing timing induced errors.

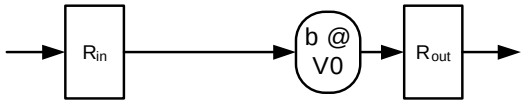
In this work we further illustrate DVAS's analogy to Dynamic Voltage Frequency Scaling (DVFS) and we show DVAS is an efficient method for scalable approximate computing due to its at wide at run-time adaptable energy-accuracy trade-off and its similar or better energy gains at iso-quality compared to other techniques.

The main novelty of this work however, lies in the generalization of the DVAS concept to pipelined systems. Pipelined systems can be made DVAS-compatible through minor changes with limited area and energy overhead at full precision. This makes major energy savings possible in high-speed pipelined systems as well. A DVAS-based pipelined multiply-accumulate used in a JPEG compression algorithm consumes 69% less than its nominal dissipation at a $2dB$ Peak-Signal-to-Noise Ratio (PSNR) drop after JPEG decompression. Energy consumption is 83% lower at a $10dB$ PSNR drop. The system can dynamically switch between these working modes.

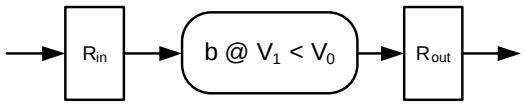
This paper is organized as follows. First, section II introduces the basic concept of Dynamic Voltage Accuracy Scaling and explains the analogy to the widespread concept of Dynamic Voltage Frequency Scaling. Second, in section III, we show how the DVAS concept can be applied to an array multiplier. We compare our approximate multiplier with the current state-of-the-art in approximate and truncated multiplication in the energy-accuracy design space. The generalization of the DVAS concept to pipelined structures is explained in section IV. Finally in section V we assess DVAS's system level accuracy and energy consumption using a JPEG compression algorithm executed with a pipelined DVAS-based multiply accumulate (MAC) system. Section VI concludes the paper.



(a) Full critical path is used at V_0



(b) A small part of the critical path is used at V_0: there is positive slack.



(c) The positive slack is minimized by reducing the supply voltage $V_1 < V_0$.

Fig. 1. The Dynamic Voltage Accuracy Scaling concept. (a) A combinational logic path existing out of part a and b has zero slack at supply voltage V_0. (b) The activity is reduced by using smaller word lengths, leading to a shorter critical path at supply voltage V_0. (c) The increased positive slack of the circuit allows for a lower supply voltage $V_1 < V_0$.

II. DVAS: Dynamic Voltage Accuracy Scaling

A. Energy savings through voltage scaling

The total power consumption P in digital systems has a dynamic and a static leakage component $P_{leakage}$ as in equation (1):

$$P = \alpha C f V^2 + P_{leakage} \tag{1}$$

The dynamic power consumption is proportional to four circuit parameters: α is the equivalent switching activity, C the capacitance, f the clock frequency and V the system's supply voltage. α depends on circuit-architecture and the application, C is mainly dependent on circuit-architecture and technology, f is determined by the critical path length, and thus by the circuit architecture and by the technology node of the design. The supply voltage V is typically fixed for a given technology node, but can be scaled to trade-off energy for performance and reliability.

Widespread technologies such as Clustered Voltage Scaling (CVS) [12] or Dynamic Voltage Frequency Scaling (DVFS) [13] exploit the quadratic dependency of digital power dissipation to V in low power designs. In CVS, certain parts of a digital system with a slack surplus will operate at a lower supply voltage than paths without slack surplus. DVFS is a system level technique, where supply voltage is dynamically decreased or increased when the system's performance requirements in terms of clock speed change. General purpose processors for example can lower their f at small workloads, when the required throughput is reduced. By doing this, slack is increased on all critical paths, allowing the supply voltage to scale accordingly. No timing errors are induced. This can result in major energy savings due to the combined effect of frequency- (linear) and voltage-scaling (quadratic) on power consumption. In analogy to DVFS, the system accuracy can be reduced in such a way that the supply voltage can be lowered as well, as further elaborated in section II-B. This again leads to an enhanced power reduction beyond DVFS, as explained in figure 1.

The theoretic extension of the DVAS concept to Dynamic Voltage Accuracy Frequency Scaling (DVAFS) is trivial, and therefore not within the scope of this text.

B. Accuracy scaling through bit width reduction

Previous work in approximate computing allows a static [3] [4] [5] trade-off between energy consumption and accuracy. Most of these works reduce energy consumption by slightly changing the logical functions of the digital arithmetic's basic building blocks. However, for many upcoming applications, system performance requirements in terms of computational accuracy strongly vary over time.

More opportunities for dynamic accuracy scaling arise when accuracy is scaled through the deliberate introduction of quantization errors at run time. This is: by reducing the number of bits used in the computation and in the representation of its value. Figure 2 plots the Root-Mean-Square Error (RMSE) and the maximal error made by multiplying with rounded input words as a function of the used bit width. Note that multiplying using 5 bit rounded words only leads to a 1.5% RMSE. The introduced errors are hence small on average, while the maximum error is deterministic and limited by the used number of bits. This is in sharp contrast to other work [3] [4], where the error probability is much lower than in our case, but the error amplitude can be large and unpredictable.

When approximating digital words by truncating or rounding them, the circuit activity α of the digital arithmetic strongly reduces. It is of course important to signal-gate the unused inputs to prevent unwanted signal propagation. Power consumption decreases linearly with α, while α itself can reduce at a higher order of the used bit width, depending on the circuit architecture. Reducing the bit width reduces activity quadratically in multipliers (see figure 3), and linearly in adder structures. The dynamic power equation can thus be rewritten as:

$$P = c_0 n^k C f V^2 \tag{2}$$

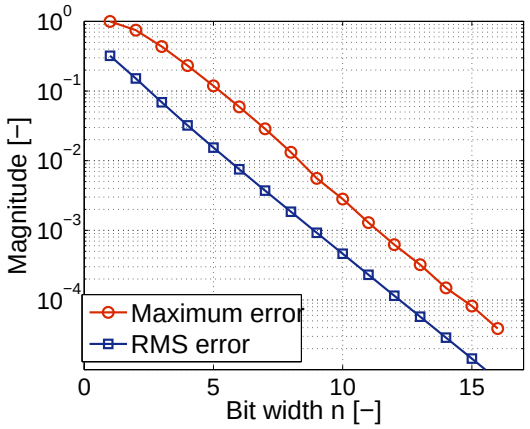


Fig. 2. Accuracy scaling through bit width reduction.

where c_0 is a constant, n is the used bit width in the computation representing the accuracy and k a power factor depending on the circuit architecture.

C. Voltage scaling through bit width reduction

By intelligently reducing the bit width of digital words, not only the activity factor α of the arithmetic circuits is reduced, but the critical paths are shortened as well. These shorter data paths lead to higher slacks on all critical paths and thus allow deeper voltage scaling, which again has a quadratic impact on power dissipation.

Figure 3 illustrates how this concept can be applied to common digital arithmetic such as an array multiplier. In the first case (red), the input words are four bit wide. As a result of this, all gates will have a nonzero switching activity and the critical path will be long, as indicated by the red arrow. In the second case (blue), only the two MSB's of each word are used. All other inputs are signal gated. The number of active gates is now reduced from 20 (red) to 6 (blue) and the critical path length is shorter, allowing voltage scaling.

The energy savings achievable in a basic DVAS system are illustrated in section III. Later, the concept is generalized to pipelined systems.

III. DVAS: Approximate Multiplier Case Study

As a first case-study, we quantify the performance of a non-pipelined signed Baugh-Wooley array multiplier architecture [14] under Dynamic Voltage Accuracy Scaling. A 16 bit implementation is simulated in a $40nm$ CMOS technology using conventional register-transfer-level (RTL) to standard-cell synthesis and simulation tools. For every accuracy setting, the critical path is derived and the operational minimum voltage maintaining the nominal operational frequency is set. The used standard-cells were characterized at two voltages: $0.9V$ and $1.1V$. Explicit simulations using these voltages are used to extrapolate the circuit performance to other supply voltages. We find these extrapolations to be accurate as long as they are outside the near- or subthreshold region, which is always the case in this work.

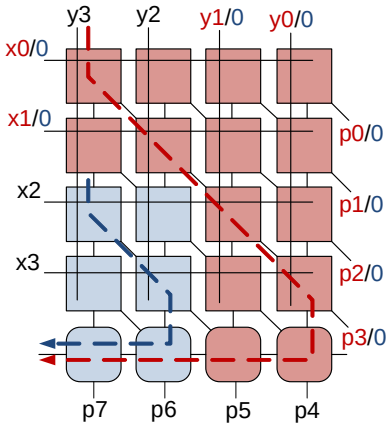


Fig. 3. An array multiplier architecture [14]. When the full bit width is used (red), the activity is high and the critical path is long. When only the 2 MSB's are used (blue), both α and the critical path length drop.

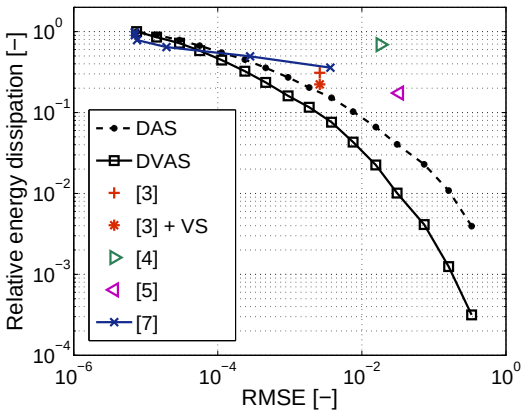


Fig. 4. Non-pipelined multiplier with continuous DVAS. Our implementation is compared to the state-of-the-art in approximate multipliers. The given energy dissipation is relative to the respective fully accurate implementation. [7] performs slightly better than DVAS at high quality, but worse at lower qualities.

Figure 4 plots the resulting dynamic energy consumption of our multiplier as a function of its mean output accuracy, described in terms of RMSE. This energy-accuracy curve is made both for the Dynamic Accuracy Scaling (DAS, where V is kept at its nominal value) and for the Dynamic Voltage Accuracy Scaling (DVAS) case. Note that DAS already strongly reduces energy consumption by reducing the activity factor. In DVAS, the energy consumption is reduced more dramatically. For example, the 8-bit case (RMSE = 0.2%) only has 60% of the critical path of the 16-bit case. This results in a $5\times$ energy drop for DAS and a total $\times 11.7$ energy drop for DVAS compared to the accurate 16-bit case. This result is striking, taking into account that the resulting deviation is only 0.2% RMSE.

The performance of our approximate multiplier under DVAS should be compared to the state-of-the-art in approximate and truncated multipliers. Figure 4 superimposes the normalized performance gains of selected recent approximate and truncated multipliers and compares them to the gains of the DVAS multiplier. References [3] [4] and [5] are static, [7] is dynamic. [3] is faster than its fully accurate alternative and can thus also be voltage scaled (VS). [7] is more energy efficient than DVAS at high accuracy, but less at lower accuracies. Figure 4 shows the maximum potential of the DVAS technique with continuous voltage scaling. In a real system, supply voltages cannot be scaled continuously, since only a discrete set of supply voltages will be available.

Dynamic Voltage Accuracy Scaling is trivial in non-pipelined circuits, but it is not in pipelined systems, which require some modifications. We generalize the DVAS-concept to pipelined systems in the next section (section IV).

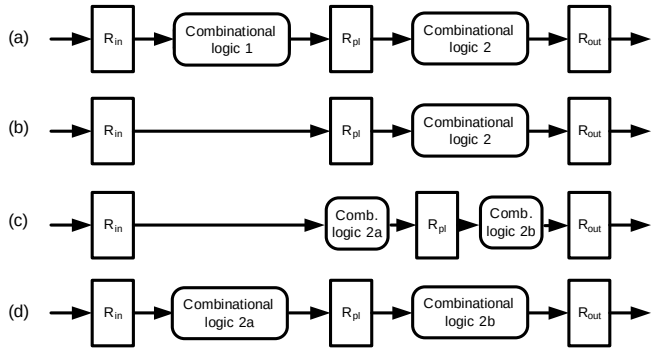
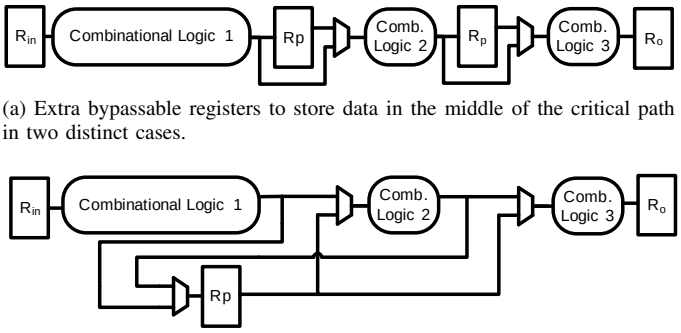


Fig. 5. The Dynamic Voltage Accuracy Scaling concept in pipelined systems. (a) A combinational logic path existing out of part 1 and 2 has zero slack at supply voltage V_0. (b) Using shorter words reduces the circuit activity but not the critical path, which is still determined by logic 2. (c) Displacing the pipeline register R_{pl} again cuts the critical path in two. Both paths now have 50% slack. (d) The supply voltage can now be lowered, resulting in zero slack at $V_1 < V_0$.



(a) Extra bypassable registers to store data in the middle of the critical path in two distinct cases.

(b) No extra registers are added, but the correct data is rewired by muxes to the pipeline registers.

Fig. 6. Two solutions to allow DVAS in pipelined systems. (a) add extra registers. (b) rewire data to the pipeline register. In the high accuracy case combinational logic 1 is active in the first cycle, and combinational logic 2 and 3 are active in the second cycle. In the low accuracy case, logic 2 is active in the first cycle and logic 3 in the second.

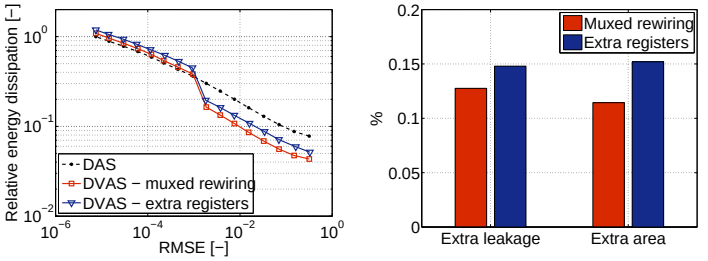
IV. DVAS: GENERALIZATION TO PIPELINED CIRCUITS

Without adaptations to a pipelined circuit DVAS cannot be applied directly, as in these systems a reduction of the used bit width does not automatically lead to a shorter critical path. This section elaborates on the generalization of the DVAS concept tot pipelined circuits.

A. General pipelined DVAS

Imagine using only half the number of bits in a two-stage pipelined digital system, such as the one in figure 5(a-b). This reduces the circuit activity in the same way as in the case without pipelining, but the critical path often does not become shorter, as one of the pipeline stages remains untouched. Pipelined circuits hence do not allow DVAS without adaptations. In a two-stage pipeline, the register optimally is always placed in the middle of the data path, balancing the delay in both pipeline stages and hence maximizing voltage scaling capabilities. Reducing the activity through smaller bit widths typically does not fulfill this requirement. It is therefore necessary to be able to dynamically displace pipeline registers upon accuracy scaling in order to benefit from the DVAS concept. To enable efficient DVAS, we propose to place bypassable pipeline registers at strategic places of the data path. Namely, at positions that will become the optimal pipeline register allocation after accuracy scaling. These latchable positions (LP) are only created for a selected set of accuracy settings, in order to keep the introduced area, energy and delay overhead low. In figure 5(c-d) for example, the pipeline register is displaced to the middle of the data path of combinational logic 2. This creates positive slack, allowing a lower supply voltage $V_1 < V_0$.

Figures 6 and 7 compare two solutions to include bypassable registers in a digital design. First, extra bypassable registers can be placed in the correct spot, as in figure 6a. The output of these registers is either the input or the clocked output, depending on a selection signal. In this way only the wanted registers can be activated. By always choosing the



(a) Dynamic energy

(b) Static energy and area overhead

Fig. 7. Comparison between two re-pipelining solutions for a 16-bit DVAS multiplier. (a) Dynamic energy dissipation. (b) Area and leakage overhead at full accuracy.

register in the middle of the current critical path, DVAS can be maximally exploited. If the registers are not used, they are clock gated to save power.

Figure 6b illustrates a second solution: muxed rewiring. Instead of placing extra registers, the appropriate signals are all routed to one single register. The register then clocks only one of the input signals, depending on a selection signal. This solution does not require extra registers as it reuses a common pipeline register, resulting in a lower energy overhead. Furthermore, the clock load will not be increased. The multiplexers still form an area and delay penalty, but this overhead is limited. Figure 7 gives a quantitative comparison between the two techniques. The second solution, 'muxed rewiring', has a lower area- and leakage overhead at the maximal accuracy compared to the standard non-DVAS multiplier, as figure 7b illustrates. Furthermore, figure 7a indicates it has lower active energy dissipation.

The number of extra registers or the added complexity due to muxes is determined by the wanted voltage-accuracy scaling granularity. To limit this overhead, only a selected set of LP will be inserted, covering a strategic subset of accuracy modes. In general, the total number of LP's or the equivalent multiplexing complexity is determined by the

following formula:

$$LP = (accuracy_{modes}) \times (S_{pipe} - 1) \quad (3)$$

In this equation S_{pipe} is the number of pipeline stages and $accuracy_{modes}$ is the number of voltage-accuracy scaling modes. For example, if two voltage-accuracy scaling modes are wanted in a two-stage pipelined system, two LP's are needed. This is one more than the non-DVAS implementation. Note that equation (3) gives an upper limit. In some cases LP's can be reused in different accuracy modes. Table I illustrates this for a three stage pipelined system with four voltage-accuracy scaling modes. This means it can discretely scale the supply voltage when 100%, 75%, 50% and 25% of the critical path is used. In this case $S_{pipe} = 2$ and $accuracy_{modes} = 4$, resulting in $LP = 8$. It is clear from table I however, that several LP's overlap, namely the ones at $2/12$ and $4/12$ of the full datapath. The needed number of LP's is thus reduced to $LP = 6$.

In general, in pipelined systems, the DVAS overhead increases quickly with the number of pipeline-stages and latchable positions. This is simulated and quantified in the next section.

B. Pipelined multipliers in the energy-accuracy space

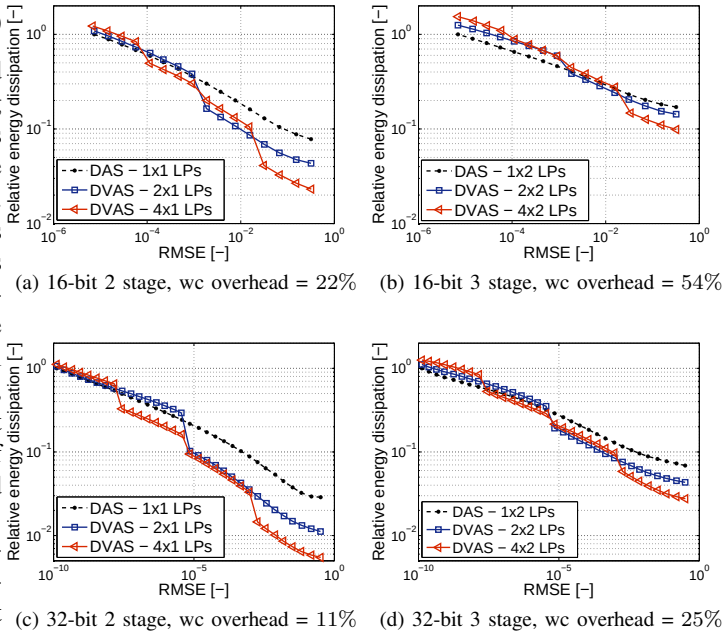
Figure 8 shows the energy-accuracy curves for 16- and 32-bit array DVAS multipliers with two and three pipeline stages simulated in $40nm$ CMOS technology.

It is clear that the 16-bit DVAS implementation is only beneficial for shallow pipeline depths. The implementation with 2×1 LP's in a two stage pipeline has an 8% dynamic energy overhead at full accuracy and achieves 91% energy savings compared to the standard implementation at 1.5% RMSE. Using 4×1 latchable positions the overhead is 22%, but it allows a wider energy trade-off up to a maximum of 98% energy savings. The 16-bit three stage DVAS multiplier, however, has a much higher overhead, 25% and 54% respectively for the implementations with 2×2 and 4×2 LP's. This overhead is too high and can only be overcome at very low accuracies. In this case Dynamic Accuracy Scaling (without extra LP's and no voltage scaling) is the best option.

Larger data paths, such as the 32-bit multiplier, benefit from pipelined DVAS in both shallow and less shallow pipelines, as the overhead due to the extra LP's is much lower in this case. The two stage pipeline with 2×1 latchable positions has a 6% dynamic energy overhead at full accuracy, but can cut 98% of the energy dissipation at an RMSE of 1.5%. The 4×1 case has an overhead of 11% but can scale down even further in energy consumption. The 32-bit three stage pipeline is now beneficial as well, because of the lower relative overheads.

TABLE I
LATCHABLE POSITIONS IN A THREE STAGE PIPELINED 16-BIT MULTIPLIER WITH FOUR VOLTAGE-ACCURACY SCALING MODES

% of critical path	100	75	50	25
MSB's used	16	12	8	4
$2nd$ LP @ $\times/12$ of 16-bit critical path	8	6	4	2
$1st$ LP @ $\times/12$ of 16-bit critical path	4	3	2	1



(a) 16-bit 2 stage, wc overhead = 22% (b) 16-bit 3 stage, wc overhead = 54%

(c) 32-bit 2 stage, wc overhead = 11% (d) 32-bit 3 stage, wc overhead = 25%

Fig. 8. Comparison of two- and three staged pipelined 16-bit multipliers under DAS and DVAS with two and four supply voltages. A three-stage multiplier

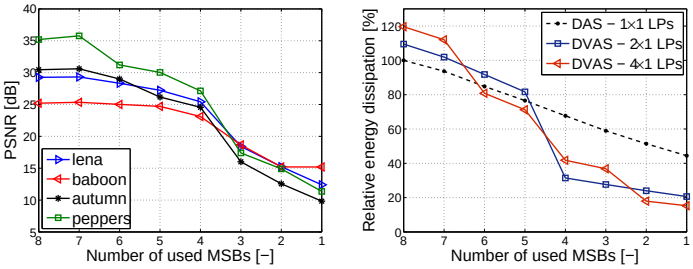
V. System level DVAS

The previous sections discussed the dynamic energy-accuracy trade-off in digital systems in general and in a pipelined array multiplier in particular. We have presented energy-accuracy curves for these multipliers. These approximate multipliers should however be tested in a larger system. We show the impact of bit width accuracy scaling in a JPEG compression algorithm, both visually as well as in the energy-accuracy design space.

A. Approximate JPEG compression

To illustrate the potential of DVAS-based circuitry in visual applications, we have implemented a DVAS-capable multiply-accumulate (MAC) unit in $40nm$ CMOS. This MAC exists out of an 8-bit two stage pipelined DVAS multiplier with two or four voltage-accuracy scaling modes and a 16-bit adder. It is used for the DCT computation of images in the JPEG compression algorithm. The input data to the DCT is approximated by rounding it to its n most significant bits (MSB). The MAC-system further consists out of a MAC-controller, the rounding arithmetic and the in- and output registers. All these are included in the energy simulations.

Figure 9a plots the accuracy of the full algorithm in terms of peak signal-to-noise ratio (PSNR) versus the used number of bits. Figure 9b illustrates the energy-accuracy trade-off for the used MAC-system for different implementations. Note that the energy gains for the 8-bit MAC are less spectacular than in the 16- and 32-bit implementations of sections III and IV, but



(a) PSNR - bit width (b) Energy - bit width

Fig. 9. (a)PSNR-performance of JPEG algorithm using an 8-bit two stage pipelined DVAS-based MAC with two or four voltage-accuracy modes for its DCT computation. (b) Relative energy dissipation of the MAC-system.

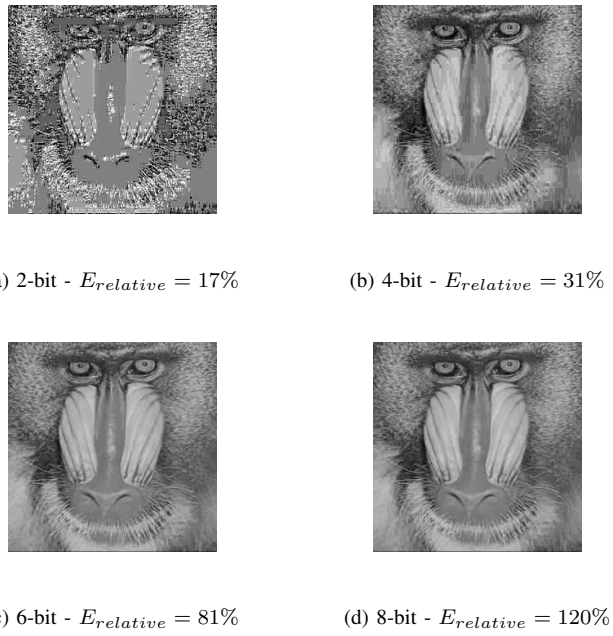
nevertheless significant. This is due to two reasons: the used DVAS data path is small (8-bit instead of 16- or 32-bit) and there is some extra overhead due to the separate adder stage, the rounding arithmetic and the extra in- and output registers.

Figure 10 visually illustrates the results of the JPEG algorithm after ideal decompression for 2-, 4-, 6- and 8-bit implementations of the MAC. The visual degradation is limited down to the 4-bit implementation (figure 10b), while the relative energy consumption of the MAC is reduced by $59-69\%$ compared to the consumption of the standard MAC at 8-bit. At 2-bit the visual degradation is more severe. However, most important features, such as the eyes and the nose of the baboon, are still retained, while the 4×1 MAC dissipates 83% less energy than the standard version. Note that a 16-bit multiplier can also be used for these bit widths if it is dynamically controlled.

VI. Conclusion

Upcoming wearable applications suffer from short battery autonomy. Several applications in recognition, mining and synthesis (RMS) and in machine-to-human or human-to-machine interactions are inherently fault tolerant and allow approximating results in return for additional energy savings. An energy-accuracy trade-off in digital arithmetic can be installed through Dynamic Voltage Accuracy Scaling, a technique analogous to Dynamic Voltage Frequency Scaling. The DVAS concept, trivial in non-pipelined circuits, is expanded to general pipelined systems. In this case, extra latchable positions should be placed in the data path. This paper carefully assessed the best strategy for implementing these latchable positions. It furthermore quantified the potential energy savings from pipelined DVAS both at data path level and on a small system level. Assessment on DVAS array multipliers validated that shallow pipelines (two stages) with many accuracy modes can be supported with limited ($< 10 - 20\%$) overhead, resulting in significant potential energy savings up to 90% or more for $< 2\%$ RMS error. In deeper pipelines (three stages and up), DVAS comes at a higher cost, which can only be justified in larger data paths such as 32-bit multipliers.

DVAS applied to a JPEG image encoder allows large system gains, without noticeable impact to user or application.



(a) 2-bit - $E_{relative} = 17\%$ (b) 4-bit - $E_{relative} = 31\%$

(c) 6-bit - $E_{relative} = 81\%$ (d) 8-bit - $E_{relative} = 120\%$

Fig. 10. Decompressed output of JPEG baboon using an 8-bit two stage pipelined DVAS-based MAC with four voltage-accuracy scaling modes.

References

[1] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," *European Test Symposium (ETS)*, 2013.

[2] S. Chippa, Vinay Venkataramani, S. T. Chakradhar, K. Roy, and A. Raghunathan, "Approximate computing: an integrated hardware approach," *Asilomar Conference on Signals, systems and computers*, 2013.

[3] C. Liu, J. Han, and F. Lombardi, "A low-power, high performance approximate multiplier with configurable partial error recovery," *Design, Automation and Test in Europe (DATE)*, 2014.

[4] P. Kulkarni, P. Gupta, and M. Ercegovac, "Trading accuracy for power with an underdesigned multiplier architecture," *International Conference on VLSI Design)*, 2011.

[5] K. Y. Kyaw *et al.*, "Low-power high-speed multiplier for error-tolerant application," *Electron devices and solid-state circuits (EDSSC))*, 2011.

[6] D. Mohapatra, V. K. Chippa, A. Raghunathan, and K. Roy, "Design of voltage-scalable meta-functions for approximate computing," *Design, Automation and Test in Europe (DATE)*, 2011.

[7] M. de la Guia Solaz, W. Han, and R. Conway, "A flexible low power dsp with a programmable truncated multiplier," *Transactions on Circuits and Systems I (TCAS-I)*, 2012.

[8] M. de la Guia Solaz and R. Conway, "Razor based programmable truncated multiply and accumulate, energy reduction for efficient digital signal processing," *Transactions on VLSI systems*, 2014.

[9] R. Likamwa *et al.*, "Draining our glass: an energy and heat characterization of google glass," *APSYS*, 2014.

[10] S. Venkataramani *et al.*, "Quality programmable vector processors for approximate computing," *International Symposium on Microarchitecture (MICRO)*, 2013.

[11] R. Hedge and N. Shanbhag, "Energy-efficient signal processing via algorithmic noise-tolerance," *International symposium on Low Power Electronics and Design*, 1999.

[12] K. Usami and M. Horowitz, "Clustered voltage scaling technique for low-power design," *International symposium on low power design (ISLPED)*, 1995.

[13] K. Choi *et al.*, "Frame-based dynamic voltage and frequency scaling for a mpeg decoder," *International conference on computer aided design (ICCAD)*, 2002.

[14] C. R. Baugh and B. Wooley, "A two's complement parallel array multiplication algorithm," *IEEE transactions on computers)*, 1973.