Energy-Quality Scalable Adaptive VLSI Circuits and Systems beyond Approximate Computing

Massimo Alioto
ECE Dept., National University of Singapore, Singapore,
Email: *massimo.alioto@nus.edu.sg*

***Abstract*— In this paper, the concept of energy-quality (EQ) scalable systems is introduced and explored, as novel design dimension to scale down energy in integrated systems for the Internet of Things (IoT). EQ-scalable systems explicitly trade off energy and quality at different levels of abstraction ("vertically"), and sub-systems ("horizontally"), creating new opportunities to improve energy efficiency for a given task and expected "quality".**

The concept of quality slack, a taxonomy of techniques to trade off energy and quality and a general EQ-scalable architecture are presented. The generality of the EQ-scaling concept is shown through several examples, ranging from logic to analog circuits, to memories and Analog-Digital Converters. Challenges, opportunities and expected energy gains are discussed to gain an understanding of the potential of the EQ-scalable integrated circuits and systems. As a result, EQ scalable systems are expected to substantially improve the energy efficiency of systems for IoT, compensating the limited energy gains that will be offered by technology and voltage scaling.

Keywords*— *VLSI design, energy efficiency, energy-quality tradeoff, adaptive energy mitigation.

I. Introduction

Energy is well known to be a major bottleneck in Systems on Chip (SoCs) for the Internet of Things (IoT) [1]-[3]. Indeed, the energy consumption per task dictates the battery lifetime, the system size and its cost, and is hence a very critical design goal [2]. Historically, Moore's law has driven down the energy consumption per task of SoCs, being responsible for 90% of the ~100X energy reduction per decade achieved since the beginning of Moore's law [4]. However, in the foreseeable future and down to the very end of its course, Moore's law is expected to contribute much less to energy (e.g., 4X [5]), given the small number of remaining CMOS generations and being the energy gain lower than 20% per generation even under the most optimistic predictions [6]. In addition, since no cost reduction per transistor is being achieved below 28nm [7]-[9], further technology scaling is not even an option in rapidly growing and cost-sensitive IoT applications. In other words, technology scaling has already delivered most of its benefits in terms of SoC energy efficiency, and now innovation needs to come from other design dimensions.

Similarly, dynamic voltage scaling has provided most of its potential energy benefits since its inception, and not much room for further voltage reductions is expected, once the supply voltage approaches the transistor threshold voltage [10]. For example, voltage scaling down to near-threshold voltages (e.g., 0.6 V) is now mainstream in 16-28nm CMOS processes [11], with very limited prospective reductions since the threshold voltage will remain essentially constant [4]. Analogously, parallelism and many-core processing is running out of steam in the vast portion of SoCs whose workload is not naturally parallelizable. For example, the number of simultaneously active cores in smartphone platforms has essentially plateaued, and the number of cores is scaling up only due to the addition of core versions covering a wider range of energy-performance tradeoffs (e.g., Tri-cluster architectures [12]). From these considerations, maintaining the historical ~100X energy reduction per decade represents a daunting challenge, under the 4X benefit brought by technology scaling. In the next decade, the remaining ~25X energy reduction will need to be achieved through further innovation, and by introducing new design dimensions and tradeoffs that further favor energy reductions.

In this paper, we introduce the broad concept of adaptive energy-quality scaling as a further design dimension to minimize energy. In general, the "quality" of a SoC component or system represents the ability to deliver results with high fidelity to the expected result. Energy-quality scaling arises from the intuitive idea that real-world systems achieve higher quality at the expense of higher energy per task, as summarized in Fig. 1. More interestingly, opportunities to reduce the energy become available when the application can tolerate lower quality at a given point of time.

In practical cases, the quality is defined by the sub-system and the application under consideration. The concept is exemplified in Table I for various sub-systems commonly encountered in SoCs. For example, conventional approximate digital circuits can lower the precision to save energy, at the expense of quality [15] (although typically in a non-adaptive manner). As another example, an analog filter or amplifier can reduce its consumption by reducing its bias current, although at the cost of poorer noise performance [16]. The energy per classification of a machine learning-based accelerator can be reduced by reducing the number of features processed, at the cost of worse misclassification rate (at least, if no overfitting is occurring [14]). The impact on quality is quantified by application-specific quality metrics, such as

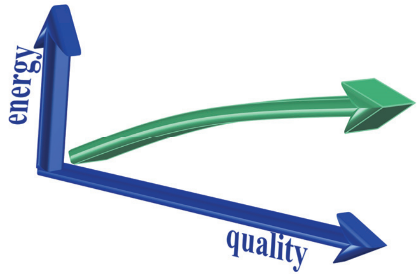


Fig. 1. General trend of energy-quality tradeoff in Systems on Chip.

978-3-9815370-8-6/17/$31.00 ©2017 IEEE

TABLE I. EXAMPLES OF ENERGY-QUALITY (EQ) TRADEOFFS IN SEVERAL SUB-SYSTEMS OF TODAY'S SYSTEMS ON CHIP FOR IOT (HIGHER QUALITY ALWAYS REQUIRES MORE ENERGY)

sub-system	knob	quality (Q) metric	EQ tradeoff when knob is increased	
			E	Q metric
digital	arithmetic precision	peak signal-to-noise ratio* (PSNR) [13]	↑	↑
analog	bias current	input-referred noise**	↑	↓
ADC	resolution	effective no. of bits*	↑	↑
SRAM	supply voltage	failure rate	↑	↓
Network on Chip	voltage swing	bit error rate**	↑	↓
switched-cap voltage converter	switching frequency	voltage ripple**	↑	↓
classification algorithm (accelerator)	no. of features***	misclassification rate**	↑	↓

* Higher values mean better quality

** Higher values mean worse quality

*** Assumption: number of feature is not excessively high, so that overfitting is not taking place [14]

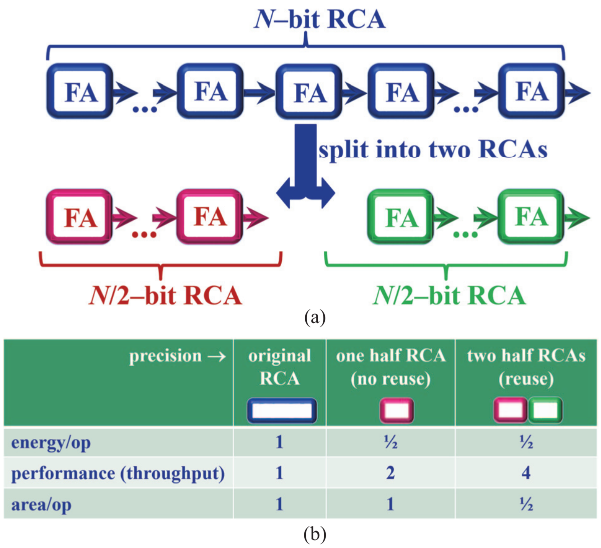
PSNR for images and videos, bit error rate for communication, or SNR for analog interfaces (see [17] for more details).

In the following, the fundamental concepts and tradeoffs in EQ-scalable circuits and systems are introduced in a unitary framework, highlighting their unique properties, and providing a taxonomy. Examples are discussed to develop a quantitative understanding of the potential of EQ scaling in typical SoC components, and envision how to achieve large energy reductions. Analogies between EQ-scalable systems and conventional VLSI designs are introduced to highlight similarities and differences.

In the remainder of this paper, Section II introduces definitions and basic concepts on EQ scaling, whereas applications and architectures are discussed in Section III. A taxonomy of techniques enabling EQ scaling is presented in Section IV, along with the analysis of their potential energy gains. Conclusions are drawn in Section VII.

II. ENERGY-QUALITY TRADEOFF: BASIC PRINCIPLES, DISTINCTIVE PROPERTIES AND ANALOGIES

As motivating example, let us consider a very simple digital block such as a Ripple Carry Adder (RCA) [18], as depicted in Fig. 2a. This RCA comprises N cascaded full adders to perform an N-bit addition, and has a delay proportional to N due to the carry propagation from the first to the last full adder [18]. Same consideration holds for the energy per operation and area per operation. Accordingly, as summarized in Fig. 2b, the energy per operation is halved when the precision is reduced from N down to $N/2$, when not reusing the remaining $N/2$ bits. The performance is improved by 2X since the delay is halved and hence the $N/2$-bit RCA can perform twice the number of operations per second. Also, the area per operation of the halved RCA is the same as the original RCA, and the quality is degraded due to the reduction in the precision. From this simple example, EQ scaling is interestingly able to simultaneously improve both energy and performance, although they are typically opposite requirements in conventional VLSI designs.



precision →	original RCA	one half RCA (no reuse)	two half RCAs (reuse)
energy/op	1	½	½
performance (throughput)	1	2	4
area/op	1	1	½

Fig. 2. a) Simple example of EQ-scalable N-bit Ripple Carry Adder (RCA) adjustable to reduce precision down to $N/2$ bits, and b) resulting energy of N-bit RCA, single $N/2$-bit RCA (i.e., the second half RCA is not reused, being simply shut down to save energy and speed up addition in the first one), and two simultaneously active N-bit RCAs.

An even more interesting observation can be made when reusing the remaining $N/2$ bits as another RCA that operates simultaneously with the first one. In this case, the performance is quadrupled due to the parallel operation of the two $N/2$-bit RCAs, and the area per operation is also halved. Hence, in this simple example EQ scaling can improve performance, energy and area efficiency all at the same time, as opposed to conventional design scenarios (e.g., voltage scaling, sizing).

From the above example, EQ scaling breaks traditional VLSI design tradeoffs, thus making design closure easier and enabling capabilities that would normally not be allowed. When the quality is explicitly treated as a knob, it typically exhibits some slack compared to the actual minimum level of quality that would be strictly required by 1) the silicon implementation, 2) the task at hand, 3) the usage scenario, and 4) the specific input dataset (physical or digital), as summarized in Fig. 3. This quality slack is analogous to the conventional timing slack in VLSI design, and its suppression can be used to improve the design in terms of energy, performance and area.

The quality slack between the very minimum ("just-enough") quality and the quality available in a SoC implementation is due to various contributions:

- RESILIENCY: one contribution is due to the energy cost of eliminating hardware faults and ensure resilient operation (e.g., timing failures in critical paths, incomplete write in an SRAM bitcell, inadequate timing window for comparator resolution). Although invariably budgeted during the design, such cost is actually un-necessary in several applications that can

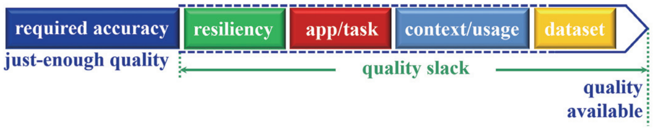


Fig. 3. Quality slack in practical design scenarios, and its main components.

tolerate occasional faults, as discussed in the Section IV. For example, overscaling the voltage and causing a few sparse pixels in a smartwatch screen to have imperfect color is not noticeable to the human eye.

- APPLICATION/TASK: different applications or tasks running on the same silicon platform might have different quality requirements. For example, an audio sub-system to perform speech recognition demands much higher quality than a voice activity detector.
- CONTEXT/USAGE: even the same application might require different levels of quality, depending on the usage scenario. For example, substantial degradation of the quality of a video stream is acceptable when a smartwatch is being used while performing physical activity, as compared to a steady user.
- DATASET: different inputs might require different levels of quality for the same objective. For example, lower quality is acceptable under highly noisy signals, as less information content can be extracted anyway.

The above quality slack contribution associated with resiliency depends on the specific die considered. All other contributions are instead independent of the die, and in general are time-varying. As a consequence, the complete elimination of the quality slack to improve energy, performance and area requires energy-quality scalable systems to adapt to both the specific chip (i.e., process/voltage/temperature variations) and to the task at hand, its context and the specific dataset. In general, partial adaptation to the above contributions is a viable design option, although it only partially reduces the quality slack.

In regard to the resiliency assurance and its cost, the above requirement of sensing and adapting to PVT variations offers an interesting analogy to conventional Error Detection And Correction (EDAC), such as Razor [19], Razor II [20], EDS [21], and the recent ultra-lightweight schemes such as iRazor [22] and RazorSRAM [23]. EDAC methods sense the timing margin at run time by detecting timing failures *in situ*, so that the system can be tuned to operate at nearly-zero design margin (i.e., the clock cycle is kept at the very minimum, without additional margin). This is crucial in IoT systems, as they tend to operate at very low voltages, and are hence subject to wide variations and design margins [2]. As summarized in Table II, traditional margined designs and EDAC methods totally suppress timing errors, although in different ways: the former ones add enough cycle time margin to absorb delay variations, whereas the latter ones operate at zero margin thanks to their ability to detect and correct timing errors. As shown in Table II, several applications can actually tolerate errors (see next section), if they are bounded in amplitude and reasonably infrequent. Such designs have a negative design margin, as they experience failures as opposed to the previous schemes. Designs that can tolerate errors but are not able to detect errors are classified as faulty, and they are associated with unusable chips (i.e., discarded at testing time). Indeed, their errors are out of control and unbounded, hence the related chips cannot be verified to be suited for a given application (i.e., its associated error bounds). On the other hand, energy-quality scalable systems need to sense and correct errors, similarly to EDAC methods, although they are allowed to experience errors even after correction. In other words, EQ-scalable designs are the natural extension and generalization of EDAC methods to error-tolerant applications.

TABLE II. Classification of VLSI designs Based on Ability to Detect/Correct Errors and Tolerate Errors

application tolerates errors ↓	error detection/correction capability	
	NO	YES
NO	traditional margined design (margin > 0)	EDAC methods (margin ~ 0)
YES	faulty design (margin < 0)	EQ scalable design (margin < 0)

In spite of the above commonalities, EQ-scalable and EDAC systems exhibit some fundamental differences, as the former ones simply need to keep errors within bounds, whereas the latter ones need to assure full error suppression. In other words, the error detection of EQ-scalable systems does not need to have full error coverage, as opposed to EDACs. This drastically reduces the traditionally large area/energy overhead of error detection in EDACs [19]-[21], making error detection circuitry much simpler in EQ-scalable systems (i.e., more area and energy efficient). As second fundamental simplification, EQ-scalable designs need to control the long-term average of errors, rather than the instantaneous occurrence of individual errors (being the quality determined by aggregate errors, rather than individual ones). This leads to a drastic simplification of the the circuitry for error correction, compared to traditional EDACs, which instead need fast correction to avoid corrupting the architectural state (e.g., before an instruction is committed in a processor). In summary, error detection/correction is still needed in EQ-scalable systems, but it can be made much slower and sparser (i.e., much lower area and energy) compared to EDACs.

III. Applications and EQ Control Architecture

Errors and degraded quality are tolerable in several applications. Applications that are statistical in nature are intrinsically robust against quality degradation and occasional errors, as the output is determined by the aggregation of a wide dataset, rather than by individual sensing/processing steps. The same applies to systems executing forms of soft computing (e.g., bio-inspired, machine learning, swarm intelligence), as their output depends on collective data, rather than individual data points.

Bounded quality degradation is also tolerable in applications that involve human perception, which is imperfect anyway and tends to look at the global output (e.g., a frame), rather than individual pieces of data (e.g., a single pixel). For similar reasons, quality degradation is typically tolerable in applications involving physical signals (e.g., sensors), as the latter are inevitably noisy. Also, applications having built-in data redundancy can tolerate quality degradation, as individual incorrect data points are mitigated by the associated redundant correct points (e.g., correlation between audio samples). As well known in error-aware computing frameworks (see, e.g., [24]), a larger margin for quality degradation is offered by applications that have high noise resiliency. Intuitively, the robustness against noise makes aggressive quality degradation tolerable, thus allowing more aggressive energy reduction.

As another fundamental differentiator between applications, data-centric applications are much more suited for EQ scaling, compared to control-centric applications. This can be intuitively understood by referring to the traditional Von

Neumann architectural model, and observing that quality degradation and imprecision are tolerable only in the data path and the memory array storing the data. Instead, inaccuracies occurring in the control path or the memory address management are not allowed, since they lead to unacceptable control flow or segmentation faults. In other words, EQ scaling brings substantial benefits only in applications and systems that are dominated by data-centric tasks.

In general, the energy reduction enabled by EQ-scalable designs are weighed by the portion of circuits that truly benefit from it (e.g., data-centric). This is analogous to performance improvements in computer architectures in which the benefit from a given speed-up technique is observed only in a portion of the architecture, as quantified by the popular Amdahl's law [25]. Similarly, the net system energy improvement through EQ scaling (i.e., $energy\ reduction$ in (1)) can be expressed in the form of Amdahl's law

$$energy\ reduction\ = \frac{1}{X\cdot\frac{E_{Q_{max}}}{E_Q}+(1-X)} \tag{1}$$

where $E_{Q_{max}}/E_Q$ is the energy saving in the EQ-scalable portion of the system when reducing the quality from its maximum Q_{max} down to the target Q, and X is the fraction of the system energy that is EQ-scalable. In (1), the application defines $E_{Q_{max}}/E_Q$ as dictated by its intrinsic noise resiliency (since the latter dictates the energy benefit of EQ scaling), and how impactful the related energy scaling is at the system level through parameter X. Applications that can effectively leverage EQ scaling are both noise resilient (i.e., $E_{Q_{max}}/E_Q \gg 1$) and data-centric ($X \approx 1$). From the plot of (1) in Fig. 4, the energy improvements enabled by EQ scaling translate in true system energy improvements only if the EQ scalable portion of the system is rather close to 100% (within a few percentage points).

In the plethora of available and prospective EQ-scalable applications, a few representative examples include [2]:

- natural IoT human-machine interfaces, such as gesture, face and speech recognition. Indeed, such devices perform tasks that are data intensive (rather than control), they are statistical in nature and typically based on soft algorithms
- multimedia wearable systems, being data intensive, related to human perception and having built-in data redundancy
- gaming, being data intensive, based on soft computing algorithms, and again related to human perception
- smart wearables with detection/classification capabilities, being data intensive, based on soft algorithms and related to human perception
- ubiquitous IoT vision systems, being data intensive, statistical, based on soft algorithms and having built-in redundancy
- smart sensor with built-in sensemaking capabilities, being physical data-driven, and based on soft algorithms.

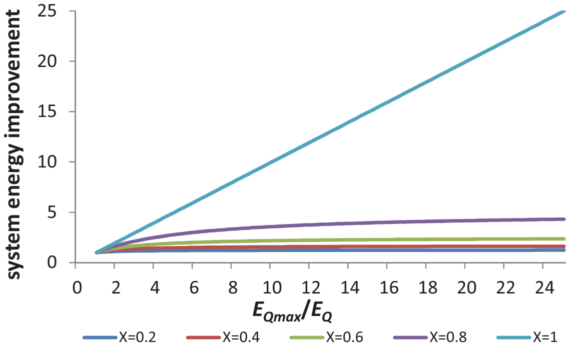


Fig. 4. System energy reduction due to $E_{Q_{max}}/E_Q$ energy improvement in E-Q scalable portion X (equivalent of Amdahl's law for EQ-scalable designs).

Let us now introduce a general architecture to control EQ scaling, which will be discussed in the context of the practical example of a smartwatch delivering video content, for the sake of simplicity. As illustrated in Fig. 5, this system can tolerate large amounts of quality degradation, as quality can be reduced when the device is operating in poor light conditions, when the user's level of attention is low, or when the device is being used during user's physical activity. Similarly, quality degradation is acceptable under low battery, to extend the device lifetime when needed. To maximize the opportunities to reduce energy, all these conditions need to be detected through sensors (e.g., light, motion sensors, battery and user activity monitor). These sensors define the context and usage scenario, and represent the input of the quality control feedback loop in Fig. 5. Through the usage model, sensor outputs are translated into the quality target (e.g., which quality is really necessary under given conditions). In turn, the quality target is used as reference for the feedback loop managed by the controller, which constantly compares the target and the actual quality measured through quality sensors. Based on the difference between the targeted and sensed quality, the controller adjusts the energy-quality knobs of on-chip EQ-scalable blocks, to make the actual quality equal to the targeted one. The architecture in Fig. 5 is very general, and each component can be implemented in silicon or in software, depending on the application constraints. For example, the controller (quality sensors) might be implemented by the Operating System or a specific on-chip controller (sub-circuits).

IV. EQ Scaling Techniques: Opportunities and Potential for Energy Reduction

Let us now introduce a general and design-centric taxonomy of methods to trade off energy and quality. For the sake of simplicity, examples of digital designs will be presented, although the concepts below are immediately generalized to all other sub-systems in Table I.

In general, the quality degradation can be adjusted through die-independent and die-dependent mechanisms, as shown in Fig. 6. By definition, die-independent EQ methods are equally effective in any considered die, and their optimization does not

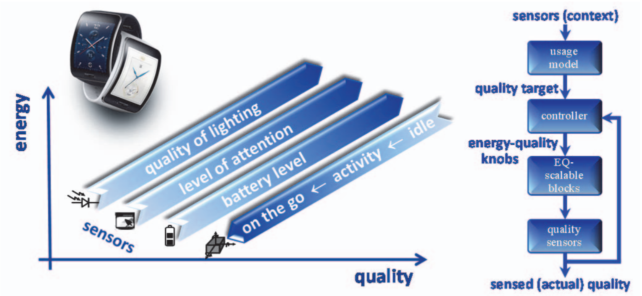


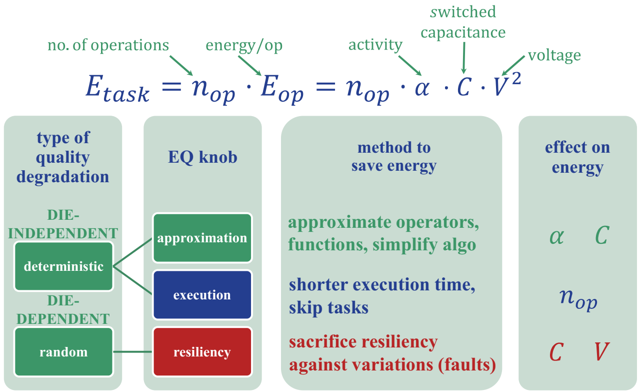
Fig. 5. Opportunities to scale down quality and energy in a smartwatch delivering video content, under poor quality of lighting, low level of user's attention, low battery level, high level of physical activity.

require any knowledge of die-specific PVT variations (i.e., these are deterministic methods). Based on the expression of the dynamic energy in digital sub-systems (see Fig. 6), a first approach to reduce the energy per task is to reduce the energy per elementary operation through algorithmic simplifications, as well as circuit- and gate-level approximations of operators and functions. For example, the wide literature on conventional approximate computing has introduced approximations in adders, multipliers, filters and Multiply-Accumulate, DCT processors, Boolean functions [26]-[34]. From a very wide survey of the literature, approximations enable energy reductions between 1.5X and 7.5X.

Beyond approximate computing similar considerations can be made in other SoC (non-digital) sub-systems. For example, the output of ADCs can be made coarser by degrading their resolution, if the latter is made tunable through circuit techniques [35], [36]. For example, in [35] we showed that proper circuit topologies enable a wide ADC resolution scalability from 3 to 11 bits, thus covering most of the typical specifications in IoT applications. Interestingly, resolution reduction by 1 bit consistently leads to approximately 2X energy reduction, which can hence be reduced by 10-30X for reasonable resolutions, compared to the full resolution.

As second class of die-independent approaches, quality and energy can be reduced by reducing the execution time, and hence reduce the number of elementary operations n_{op} per task (see Fig. 6). Such methods are simple to implement, as they are mostly applicable to iterative algorithms and tasks. As a few examples, reduced time window has been exploited in digital filtering [37] (e.g., lower FIR filter order when the filter selectivity is less stringent). Input data or vector basis subsampling is another effective approach [38]. Also, early termination of iterative refinement algorithms have been widely investigated [39], including "anytime" algorithms [40]. All these methods are well established, as they mostly deal with adaptive loop termination, which has been already explored widely in software programming. Accordingly, although very useful and effective with energy gains from a few units to 10X, the research area aiming at the reduction of the execution time is not really active.

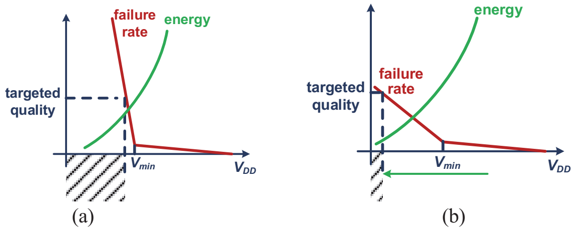
On the other hand, die-dependent EQ methods require an EQ optimization that is influenced by die-specific random variations. In other words, they explicitly manage the tradeoff between energy and random variations, and hence resiliency. As shown in the bottom-right quadrant of Table II, operation under negative design margin leads to occasional and infrequent faults occur during operation. For example, operation in this regime is achieved by overscaling the clock frequency or the supply voltage beyond the zero-margin value, thus improving the performance and/or the energy efficiency at the expense of resiliency. Most of the research in this area has been focused on logic, and energy gains of 1.4-5X have been demonstrated [41], [42]. The actual challenge to achieve such energy gains is to mitigate the very rapid increase in the failure rate at negative margins. Indeed, it is well known that operation at V_{DD} lower than the minimum voltage V_{min} ensuring perfectly correct operation leads to a failure rate that increase exponentially when reducing V_{DD} [20], [21]. This prevents any significant reduction in V_{DD} below V_{min} for reasonable quality targets, as the quality is degraded dramatically even for small reductions in V_{DD} (see Fig. 7a). Accordingly, the real challenge to achieve EQ scaling is to make the quality degradation graceful at low voltages, so that more aggressive voltage (i.e., energy) downscaling is allowed without degrading the quality excessively (see Fig. 7b).



switched capacitance
no. of operations
energy/op
activity
voltage
$E_{task} = n_{op} \cdot E_{op} = n_{op} \cdot \alpha \cdot C \cdot V^2$
type of quality degradation
EQ knob
method to save energy
effect on energy
DIE-INDEPENDENT
deterministic
DIE-DEPENDENT
random
approximation
execution
resiliency
approximate operators, functions, simplify algo
shorter execution time, skip tasks
sacrifice resiliency against variations (faults)
α C
n_{op}
C V

Fig. 6. Taxonomy of methods to trade off energy and quality.

Techniques to make quality degradation graceful have been mostly demonstrated for logic circuits [20], [21], whereas very little research has been devoted to other SoC sub-systems, such as SRAMs and Networks on Chip. Actually, the latter ones are much more critical in terms of resiliency, and define the true challenge ahead to be addressed in die-dependent EQ scaling. This is because they consist of the repetition of a large number of nominally equal circuits (i.e., bitcells, links), which vastly increases the failure rate compared to logic circuits [18]. On the other hand, such circuits exhibit a high level of regularity, and hence offer unique opportunities to manage the EQ tradeoff with minimal circuitry. For example, in [43] we have shown that graceful quality degradation can be achieved in SRAMs by introducing bit-level EQ tradeoff, via selective/non-uniform assist techniques and Error Correcting Code. This is very different from conventional designs where such techniques are applied uniformly across bit positions, and entails a very small area overhead (1%). More in detail, leveraging the higher importance of MSBs in terms of quality, energy is selectively spent to improve the resiliency of MSBs (e.g., more aggressive assist techniques are used), while saving energy in LSBs. Accordingly, the quality is improved and degraded more gracefully at low voltages, thus allowing for more aggressive voltage scaling and hence energy reduction (see Fig. 7b). In addition to the quadratic energy advantage of pure voltage scaling, the above approach further reduces the energy by more than 2X [44]. Similar results are obtained in Networks on chip.



(a)

(b)

Fig. 7. Qualitative trend of failure rate (i.e., quality) vs. V_{DD} in a) convention designs with disgraceful (exponential) quality degradation at low voltages, b) EQ scalable designs with graceful quality degradation at low voltages (more aggressive voltage scaling is allowed for given quality target).

In summary, the main challenge ahead is to enable true EQ scalability and graceful quality degradation in critical SoC subsystems (e.g., memories, NoCs, analog), well beyond the prior effort that has been spent on logic and approximate computing.

V. Conclusion

In this paper, EQ scalable systems have been introduced along with their unique properties. Interestingly, EQ scalable designs break traditional design tradeoffs, allowing simultaneous improvement of traditionally opposite requirements (e.g., area, energy, performance). The existence of quality slack and its decomposition into individual contributions have been discussed in detail. The analysis of its contributions mandates adaptation of EQ scalable systems to fluctuating operating conditions, dataset and chip-specific variations. The related area/energy cost of adapting to variations and detecting/correcting errors is much lower than traditional EDACs. As an interesting perspective, actually EQ scalable systems have been shown to be a generalization of EDACs. A general architecture of EQ scalable systems has also been illustrated.

A taxonomy of available techniques for EQ scaling has been discussed, categorizing them into three classes. Interestingly, these three classes offer orthogonal solutions, and can be used simultaneously to further reduce the energy. Thus, the overall energy reduction is simply the product of the reductions in each class of techniques. Since each class of techniques is capable of reducing energy by several units (from 1.5X to 7.5X), EQ scalable systems are likely to provide the energy reductions that Moore's law cannot provide any longer in the decade ahead.

VI. Acknowledgement

The kind support by the MOE2014-T2-1-161 grant from the Singaporean Ministry of Education is acknowledged.

References

[1] M. Alioto, Guest Editorial for the Special Issue on "Ultra-Low-Voltage VLSI Circuits and Systems for Green Computing", *IEEE Trans. on Circuits and Systems – part II*, vol. 59, no. 12, pp. 849-852, Dec. 2012.

[2] M. Alioto (Ed.), *Enabling the Internet of Things - from Integrated Circuits to Integrated Systems*, Springer, 2017.

[3] Y. Guo, Y. Fang, "Electricity Cost Saving Strategy in Data Centers by Using Energy Storage" *IEEE TPDS*, vol. 24, no. 6, pp. 1149-1160, June 2013.

[4] J. Koomey, S. Berard, M. Sanchez, H. Wong, "Implications of Historical Trends in the Electrical Efficiency of Computing" *IEEE Annals of the History of Computing*, pp. 46-54, March 2011.

[5] M. Alioto, "Designing (Relatively) Reliable Systems with (Highly) Unreliable Components" keynote at *NEWCAS 2016* conference, Vancouver (CA), June 24-26, 2016.

[6] International Technology Roadmap for Semiconductors: 2015 edition [online]. *http://www.semiconductors.org/main/2015_international_technology_roadmap_for_semiconductors_itrs, 2015*.

[7] R. Courtland, "The Status of Moore's Law: It's Complicated," IEEE Spectrum, Oct 28, 2013.

[8] H. Jones , "Why Migration to 20nm Bulk CMOS and 16/14nm FinFETs Is Not Best Approach for Semiconductor Industry" white paper, 2014.

[9] T. Simonite, "Moore's Law Is Dead. Now What?," MIT Technology Review, May 13, 2016.

[10] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits," Proceedings of the IEEE, vol. 98, no. 2, Feb. 2010.

[11] TSMC technology summary – available at *http://www.tsmc.com/english/dedicatedFoundry/technology/16nm.htm*.

[12] H. T. Mair, et Al., "A 20nm 2.5GHz ultra-low-power tri-cluster CPU subsystem with adaptive power allocation for optimal mobile SoC performance," in *IEEE ISSCC Dig. Tech. Papers*, 2016, pp. 76–77.

[13] S. Winkler, P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *IEEE TB*, vol. 54, no. 3, pp. 660-668, March 2008.

[14] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[15] S. Mittal, "A Survey of Techniques for Approximate Computing," ACM Computing Surveys, March 2016.

[16] P. Gray, R. Meyer, *Analysis and Design of Analog Integrated Circuits* (5th ed.), John Wiley & Sons, 2009.

[17] M. A. Breuer, "Intelligible Test Techniques to Support Error-Tolerance," in proc. of *ATS 2004*.

[18] N. Weste, D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective* (4th ed.), Addison-Wesley, 2011.

[19] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, T. Mudge, "Razor: a low-power pipeline based on circuit-level timing speculation," in Proc. of *MICRO-36*, pp. 7- 18, Dec. 2003.

[20] S. Das, C. Tokunaga, S. Pant, W.-H. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, D. T. Blaauw, "Razor II: In Situ Error Detection and Correction For PVT and SER Tolerance," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, Jan. 2009.

[21] K. A. Bowman, J. W. Tschanz, N. S. Kim, J. C. Lee, C. B. Wilkerson, S.-L. Lu, T. Karnik, V. De, "Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance," *IEEE J. Solid-State Circuits*, pp. 49–63, Jan. 2009.

[22] Y. Zhang, M. Khayatzadeh, K. Yang, M. Saligane, M. Alioto, D. Blaauw, D. Sylvester, "iRazor: 3-Transistor Current-Based Error Detection and Correction in an ARM Cortex-R4 Processor," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2016.

[23] M. Khayatzadeh, M. Saligane, J. Wang, M. Alioto, D. Blaauw, D. Sylvester, "A Reconfigurable Dual Port Memory with Error Detection and Correction in 28nm FDSOI," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2016, pp. 310-311.

[24] R. Hegde, N. R. Shanbhag, "Soft digital signal processing," *IEEE Trans. on VLSI Systems*, vol. 9, no. 6, pp. 813-823, June 2001.

[25] J. Hennessy, D. Patterson, *Computer Architecture – A Quantitative Approach (5th ed.)*, Morgan Kaufmann, 2012.

[26] Y. Kim, Y. Zhang, P. Li, "Energy Efficient Approximate Arithmetic for Error Resilient Neuromorphic Computing," IEEE TVLSI, vol. 23, no. 11, Nov. 2015.

[27] H. Almurib, T. Kumar, F. Lombardi, "Inexact Designs for Approximate Low Power Addition by Cell Replacement," in Proc. of *DATE 2016*.

[28] J. George, B. Marr, B.E.S. Akgul, K.V. Palem, "Probabilistic Arithmetic and Energy Efficient Embedded Signal Processing," in CASES 2006, pp. 158-168.

[29] P. Kulkarni, P. Gupta, M. Ercegovac, "Trading accuracy for power with an underdesigned multiplier architecture," in Proc. 24th Intl. Conf. on VLSI Design, pp. 346–351, January 2011.

[30] C. Liu, J. Han, F. Lombardi, "A Low-Power, High-Performance Approximate Multiplier with Configurable Partial Error Recovery," in DATE'14.

[31] V. K. Chippa, D. Mohapatra, A. Raghunathan, K. Roy, S. T. Chakradhar, "Scalable effort hardware design: Exploiting algorithmic resilience for energy efficiency," in Proc. DAC 2010, pp. 555–560, June 2010.

[32] H. Kaul, M. Anders, S. Mathew, S. Hsu, A. Agarwal, F. Sheikh, R. Krishnamurthy, S. Borkar, "A 1.45GHz 52-to-162GFLOPS/W Variable-Precision Floating-Point Fused Multiply-Add Unit with Certainty Tracking in 32nm CMOS," ISSCC 2012.

[33] A. Lingamneni, C. Enz, J.-L. Nagel, K. Palem, C. Piguet, "Energy Parsimonious Circuit Design through Probabilistic Pruning," in Proc. of *DATE'11*.

[34] S. Venkataramani, A. Sabne, V. Kozhikkottu, K. Roy, A. Raghunathan, "SALSA: Systematic Logic Synthesis of Approximate Circuits," in Proc. of DAC 2012.

[35] L. Freyman, D. Fick, M. Alioto, D. Blaauw, D. Sylvester, "A 346µm2 VCO-based, Reference-Free, Self-Timed Sensor Interface for Cubic-Millimeter Sensor Nodes in 28nm CMOS," *IEEE Journal of Solid-State Circuits*, Nov. 2014.

[36] M. Yip , A. P. Chandrakasan, "A resolution-reconfigurable 5-to-10b 0.4-to-1V power scalable SAR ADC," *ISSCC 2011*.

[37] J. T. Ludwig, et Al., "Low-Power Digital Filtering Using Approximate Processing", *IEEE JSSC* 1996

[38] Chippa, et Al., "Scalable effort hardware design - exploiting algorithmic resilience for (...)", in Proc. of *DAC 2010*

[39] S. Sidiroglou, S. Misailovic, H. Hoffmann, M. Rinard. "Managing performance vs. accuracy trade-offs with loop perforation," *ACM SIGSOFT Symp. and the 13th European Conference on Foundations of Software Engineering*, 2011.

[40] A. A. Saba, et Al., "Anytime Algorithms for GPU Architectures", *IEEE JSSC* 1996.

[41] F. Kurdahi, et Al., "F. Kurdahi, A. Eltawil, K. Yi, S. Cheng, and A. Khajeh. Low-power multimedia)...)," *IEEE TVLSI*, 2010.

[42] R. Hedge, N. Shanbhag, "Soft Digital Signal Processing," *IEEE TVLSI*, 2001.

[43] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, M. Alioto, "SRAM for Error-Tolerant Applications with Dynamic Energy-Quality Management in 28nm CMOS," *IEEE JSSC*, vol. 50, no. 3, pp. 1310-1323, March 2015.

[44] F. Frustaci, D. Blaauw, D. Sylvester, M. Alioto, "Approximate SRAMs with Dynamic Energy-Quality Management," *IEEE TVLSI*, vol. 24, no. 6, June 2016.