# Survey of Precision-Scalable Multiply-Accumulate Units for Neural-Network Processing

Vincent Camus[1,2], Christian Enz[1], Marian Verhelst[2]
[1]ICLAB, EPFL, Switzerland
[2]ESAT-MICAS, KU Leuven, Belgium
vincent.camus@epfl.ch

***Abstract*—The current trend for deep learning has come with an enormous computational need for billions of Multiply-Accumulate (MAC) operations per inference. Fortunately, reduced precision has demonstrated large benefits with low impact on accuracy, paving the way towards processing in mobile devices and IoT nodes. Precision-scalable MAC architectures optimized for neural networks have recently gained interest thanks to their subword parallel or bit-serial capabilities. Yet, it has been hard to make a fair judgment of their relative benefits as they have been implemented with different technologies and performance targets. In this work, run-time configurable MAC units from ISSCC 2017 and 2018 are implemented and compared objectively under diverse precision scenarios. All circuits are synthesized in a 28 nm commercial CMOS process with precision ranging from 2 to 8 bits. This work analyzes the impact of scalability and compares the different MAC units in terms of energy, throughput and area, aiming to understand the optimal architectures to reduce computation costs in neural-network processing.**

***Index Terms*—ASIC, deep neural networks, neural-network accelerators, multiply-accumulate, MAC, configurable circuits, precision-scalable circuits.**

## I. Introduction

The current trend for deep-learning applications, such as image classification and speech recognition, has come with an enormous computational need. Indeed, state-of-the-art Deep Neural Networks (DNN) require billions of Multiply-Accumulate (MAC) operations, the fundamental component of their convolution layers, as well as fetching of millions of network parameters (weights) and feature maps (activations). Many hardware improvements have recently been proposed, innovating at different abstraction levels to solve both memory and computational bottlenecks [1], [2].

Exploiting reduced precision has demonstrated huge benefits with no or negligible impact on the network accuracy, paving the way towards embedded DNN processing in mobile devices and IoT nodes. It has led to a new trend for precision-scalable neural processors to minimize energy at target performance without giving up flexibility. Recent papers have introduced run-time configurable MAC architectures optimized for deep learning, built either with high parallelization capabilities [3], [4] or bit-serial approaches [5], [6].

However, it is difficult to assess the efficiency of these architectures for two reasons. 1) They have been implemented using diverse process technologies, bitwidths and scalability levels, and have been integrated within quite different systems, with various memory sizes and interfaces, or targeting different trade-offs. This makes it impossible to extract the precise cost of the MAC or its contribution to the system. 2) While nearly all these works detail the relative efficiency breakdown of scaling down precision in their design, few evaluate their absolute performance against a baseline design, without configuration nor parallelization overheads.

TABLE I. Reviewed designs and their scalability features

| MAC architecture | Weight scalability | Activation scalability |
|---|---|---|
| Conventional [7] | Data gating | Data gating |
| DVAFS (Envision [3]) | Symmetric subword parallel | |
| Divide-and-conquer (DNPU [4]) | Subword parallel | Data gating |
| Bit serial (UNPU [5]) | Serial | Data gating |
| Multi-bit serial (this work) | Serial | Data gating |

In this work, the most prominent precision-scalable MAC accelerators are presented, implemented, and compared in a fair way under different precision scenarios. All circuits are synthesized in a 28 nm CMOS process with bitwidth precision ranging from 8 bits down to 2 bits using both hardware scalability features and data gating. This study analyzes the impact of precision scalability on energy efficiency, especially its unavoidable energy overhead due to the introduced configurability, and gives a global picture of energy and throughput-per-area capabilities of the different architectures.

This paper is organized as follows. Section II introduces general considerations about scaling precision and details the deducted methodology to compare the different designs with impartiality. Section III surveys the considered architectures. Finally, Section IV analyzes the energy breakdown of scalability individually and presents a comparison across all approaches.

## II. Considerations and Methodology

### A. Scalability by Design and by Data Gating

Precision-scalable MACs have gained interest following the observation that the optimum bit-width for a DNN strongly varies from one application to another, and even across the different layers of a single DNN [7].

The easiest way to scale precision in arithmetic circuits is to use data gating, i.e. reducing precision by simply zeroing operands' LSBs to avoid unnecessary toggling in the circuit. This only introduces marginal overhead in the MAC periphery, leaving the MAC untouched. In contrast, precision-scalable architectures embed additional features *by design*, inside the MAC, to increase parallelism or to clock-gate unused parts of the circuit. However, these features have a cost. First, they imply some control circuitry, outside of the MAC unit, which can fortunately be shared among an array of processing elements.

More importantly, they induce hardware overhead in each MAC, leading to increased area, delay and power consumption. This is why neural processing elements often support a limited set of precisions by design, such as 2b, 4b and 8b inputs.

Two schools of thought exist for scaling DNN computations. Research originally aimed at *weight-only* precision scaling rather than activation, mainly to decrease model sizes. But late works started to quantize both weights and activations, which can be simplified as a *symmetric* scaling scenario, where both weights and activations are scaled at the same precision.

### B. Scope and Methodology of the Study

This paper evaluates the precision-scalable MAC designs listed in Table I, most of them presented at ISSCC 2017 and 2018. They are studied for 2b, 4b and 8b precisions, these values being commonly found in literature. Both symmetric and weight-only scaling scenarios are considered.

Multiple implementations of each architecture are made, varying their *level of scalability* starting from their 8b baseline. For instance, a 1-level scalable design allows to scale from 8b down to 4b by design, the 2b mode carried out by data gating over the 4b mode. A 2-level scalable design directly allows to scale down to 4b and 2b by design. Circuits are implemented in such a way that for any precision scenario, the accumulator bitwidth is exactly 4b larger than the multiplication range.

All architectures are generated from SystemVerilog descriptions with signed-weight and unsigned-activation representations. They are built at the highest level of abstraction without individual optimization. All circuits are synthesized with identical compiler options following a multi-mode timing optimization to find the best delay trade-off for all scaling scenarios. Finally, a design-space exploration is performed across all the achievable throughputs.

For each circuit and each precision mode, Mentor Questa is used to generate VCD switching information from a timed gate-level simulation at best clock period. The simulation for each mode consists out of 10,000 operations with Gaussian-distributed random inputs. VCD files are then fed back to Cadence for power estimation. Power and area of input registers and overheads that can be shared among multiple processing elements (e.g. control logic or finite state machines) are not included in the reported results.

## III. Survey of Multiply-Accumulate Units

### A. Data-Gated Conventional MAC

The baseline of this study is a data-gated conventional MAC unit as in [7]. When scaled precision is used, only the MSBs are used for computation while the LSBs are kept at zero. Hence, the switching activity is reduced. As the critical path going only through the MSBs is shorter, the frequency can dynamically be increased or the supply voltage can be lowered at equal throughput.

This is illustrated in Fig. 1, showing from left to right the use with full precision, 4-bit symmetric scaling, and 2-bit weight-only scaling. When using scaled-precision computations, some parts of the multiplier and accumulator logic keep unused, as shown by the grey stripes. However, as this MAC does not embed any configurability feature, such as selective clock gating, all registers stay clocked despite no data reach LSBs.

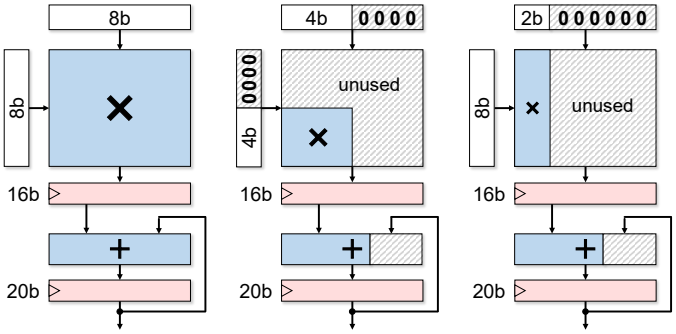


Fig. 1. Data-gated conventional MAC with an example of symmetric 4b×4b scaling (middle) or weight-only 2b×8b scaling (right).

### B. DVAFS MAC

Dynamic Voltage-Accuracy-Frequency Scaling (DVAFS) was introduced by Moons *et. al* [3]. Built on an array multiplier circuit, the DVAFS MAC reuses full-adder cells that are inactive at scaled precision. This scales together weight and activation with symmetric subword parallelism, as shown on Fig. 2. Similar to the data-gated conventional multiplier, the shortened critical path permits an increased clock frequency. Note that processing at full precision with DVAFS comes at a slight energy and area penalty due to the complex configuration and sign-compensation logic overheads, and the larger registers required for extra accumulation bits.

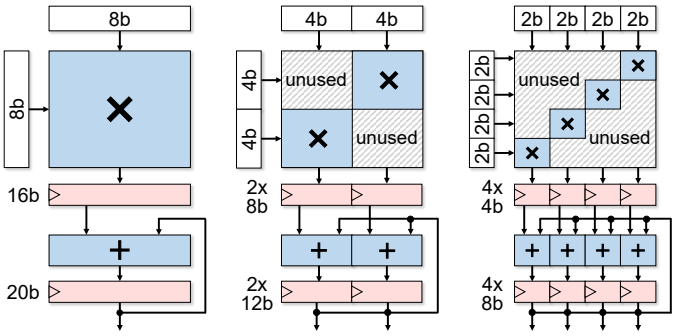


Fig. 2. Symmetric precision scaling in a DVAFS MAC configured for either one 8b×8b, two 4b×4b, or four 2b×2b operations per cycle.

### C. Divide-and-Conquer Strategy

The Deep Neural Processing Unit (DNPU), introduced by Shin *et. al* [4], uses a reconfigurable multiplier with a *divide-and-conquer* (D&C) approach on one operand. As shown on Fig. 3, the full-precision multiplier is built out of shifted binary additions of partial products. By this manner, intermediate sums directly correspond to scaled multiplication results.

Only one operand is subword parallel, doubling the number of operations per cycle for each scalability level. The other operand is common to all subword computations, restricting its use to repeated operand, but allowing parallelization when one operand has to stay at full precision. Scaling of the second operand is always possible by data gating.

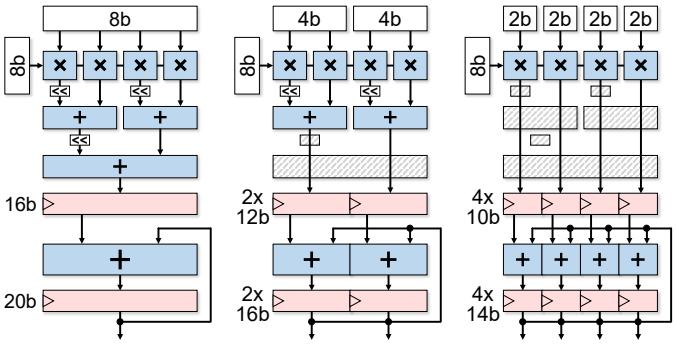


Fig. 3. Weight-only precision scaling in a D&C MAC configured for either one 8b×8b, two 4b×8b, or four 2b×8b operations per cycle.

### D. Bit-Serial Designs

Bit-serial designs have recently gained attention with both the Unified Neural Processing Unit (UNPU) by Lee *et. al* [5] and the QUEST log-quantized 3D-stacked inference engine by Ueyoshi *et. al* [6]. Indeed, bit-serial operand feeding implicitly allows fully-variable bit precision. Considered in this study, the UNPU bit-serial MAC receives weights through 1-bit iterations while activations are sent in a parallel manner, as illustrated in Fig. 4. Scaling activation is possible by data gating.

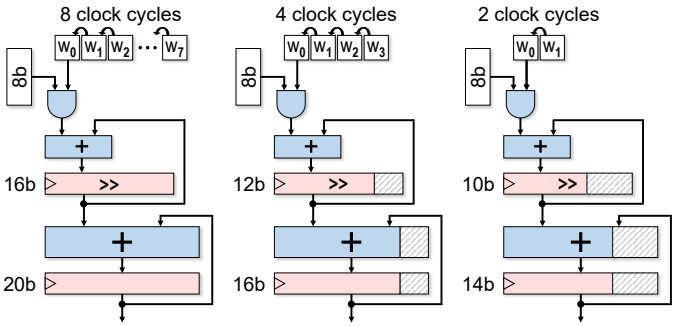


Fig. 4. Weight-only precision scaling in a bit-serial MAC configured for either 8b×8b, 4b×8b, or 2b×8b operations.

This work extends the original bit-serial concept by introducing *multi-bit serial* designs. Fig. 5 shows the example of a 4-bit serial MAC, where weights are fed 4 bits at a time. This scheme requires only 2 clock cycles for an 8-bit computation, hence reducing the energy consumed in the clock tree and registers. Lower precision can be obtained by gating the unnecessary bits. This survey includes both 2-bit and 4-bit serial MACs.

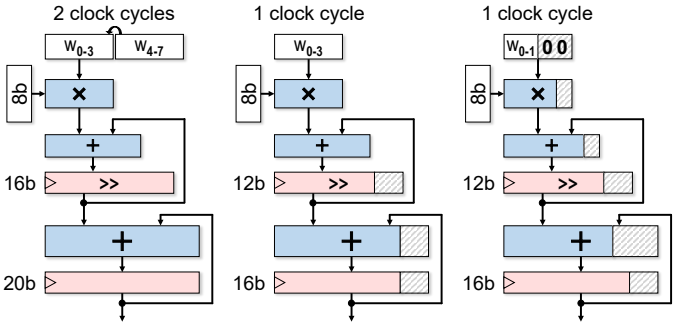


Fig. 5. Weight-only precision scaling in a 4-bit serial MAC configured for either 8b×8b, 4b×8b, or 2b×8b (by gating the 4b×8b) operations.

## IV. Results and Comparative Study

### A. Precision-Scaling Energy Breakdown

Figs. 6-8 show the breakdowns of energy per operation when scaling precision for each type of architecture, selecting the circuit with the lowest energy per operation (without consideration for throughput or area). The left subfigures show symmetric scaling scenarios while the right ones show weight-only scaling. Energy values are normalized to the full-precision data-gated conventional MAC drawn with a solid black line.

Processing at full precision with scalable designs comes with some energy penalty. 8-bit computations with DVAFS (Fig. 6) consume 13 % and 28 % more energy than data gating for embedding 1 and 2 levels of scalability, respectively. For D&C circuits (Fig. 7), these overheads are roughly similar with respectively 18 % and 26 % extra energy per 8-bit operation.

Despite their efficient use of area, serial designs (Fig. 8) require much more energy at high precisions due to their need for several clock cycles per computation, diluting the power into the clock tree and the multiplication registers. Reassuringly, the proposed multi-bit designs come at a lower energy penalty: the 4-bit serial MAC reduces the energy overhead to 57 %.

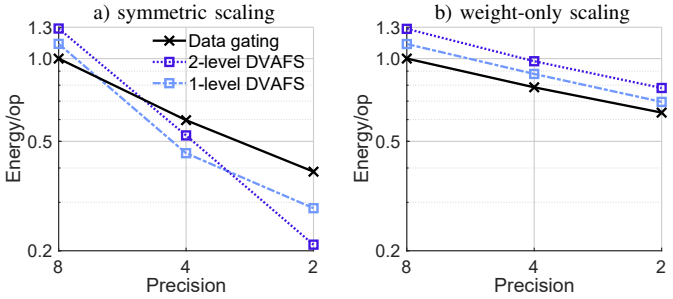


Fig. 6. Normalized energy/op with precision scaling in a DVAFS MAC.

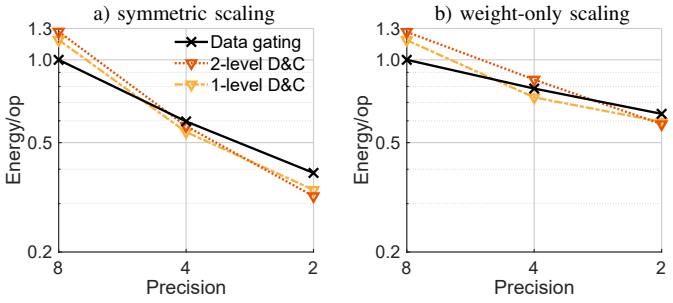


Fig. 7. Normalized energy/op with precision scaling in a D&C MAC.

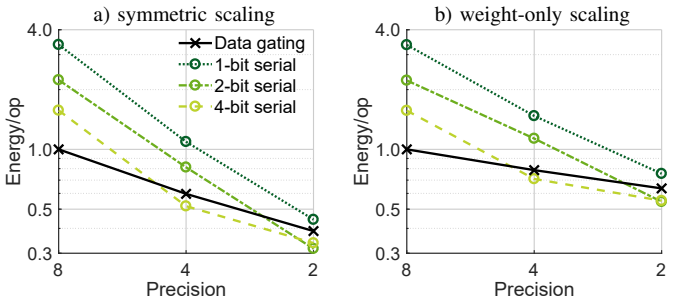


Fig. 8. Normalized energy/op with precision scaling in a serial MAC.

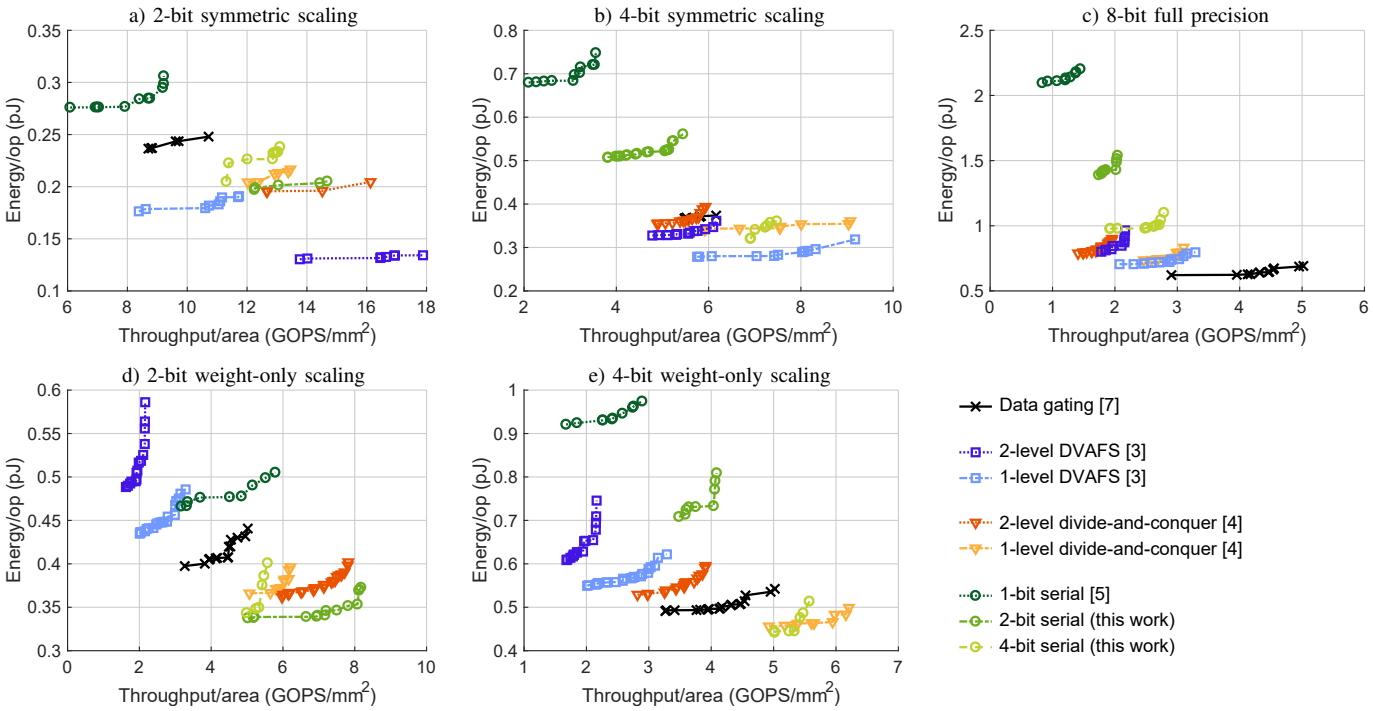


Fig. 9. Comparison of MAC architectures synthesized in a 28 nm CMOS process in terms of energy/op and throughput/area under each precision mode.

Scaling both weights and activations (left subfigures) with data gating leads to energy savings in a linear way with respect to the bit precision. In comparison, precision-scalable MACs show a steeper slope, meaning that they save energy in a superlinear way with bit precision. It is although insufficient for the 1-bit serial design to compensate its energy penalty. Below 4 bits, 1-level scalable MACs (including 4-bit serial) can only scale precision through data gating, returning to the slope of the baseline. Overall, DVAFS MACs display the best energy reductions across the entire precision range.

When preserving full-precision activations (right subfigures), savings are without exception far lower. As it is unsuitable for weight-only scaling, DVAFS designs are penalized by their overhead compared to the baseline MAC. D&C and serial designs both follow similar trends as for symmetric scaling. 1-level scalable designs appear to be the best trade-off for weight-only scaling: compared to 2-level scalable circuits, they are superior in 4 bits and almost equivalent in 2 bits.

### *B. Comparative Study*

Fig. 9 gives an overall comparison of the different MAC architectures for each precision scenario in terms of energy per operation and throughput per area. The best designs are towards the bottom-right corners of subfigures.

At full precision (Fig. 9c), data gating is undoubtedly the most efficient technique, capable of the highest throughput per area, followed first by 1-level and then by 2-level scalable designs which suffer from longer critical path.

When scaling precision symmetrically (Figs. 9a-b), the DVAFS architecture clearly outperforms all other architectures in terms of energy per operation. In this mode, its subword parallelism also yields to great throughput-area efficiency.

Note that with symmetric precision scaling, 1-level scalable circuits stay the best compromise at 4 bits and above, this trend reverses at 2-bit precision (inversion of bright and dark curves), except for 1-bit serial MACs which stay the worst trade-off due to their low speed and energy efficiency.

Interestingly, D&C and 4-bit serial MACs prove capable of good symmetric scaling, despite not being optimized for it. This is due to the optimization of the first adder stage of these designs, which benefits together to all modes, while for DVAFS, the same hardware has to be optimized for different objectives as the critical path changes from one mode to the other.

When reducing weight precision only (Figs. 9d-e), D&C and multi-bit serial architectures are the best trade-offs between energy and throughput per area. By-passing internal additions, D&C designs are advantaged for throughput, while 4-bit serial circuits slightly exceed in terms of energy per operation. 2-bit serial MACs outrun these two for 2-bit precision only.

## V. Conclusion and Future Work

This work has surveyed different precision-scalable MAC architectures, namely DVAFS, divide-and-conquer and bit serial. This later has been enhanced by introducing multi-bit serial designs. All architectures have been synthesized in a 28 nm process across a wide range of performance and precision scenarios, and compared in terms of energy and throughput per area. This preliminary study has shown that DVAFS surpasses the state of the art for symmetric scaling, while multi-bit serial and divide-and-conquer strategies exceed when scaling weights only. It has also highlighted that less scalability levels can be a good trade-off thanks to lower circuit overheads. Future works could propose a more extensive analysis and cover additional configurable or low-precision design techniques [8]–[10].

## REFERENCES


[1] M. Verhelst and B. Moons, "Embedded deep neural network processing: Algorithmic and processor techniques bring deep learning to IoT and edge devices," in *IEEE Solid-State Circuits Magazine*, vol. 9, no. 4, Nov. 2017, pp. 55–65.

[2] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," in *Proceedings of the IEEE*, vol. 105, no. 12, Dec. 2017, pp. 2295–2329.

[3] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm FDSOI," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2017, pp. 246–247.

[4] D. Shin, J. Lee, J. Lee, and H. Yoo, "DNPU: An 8.1TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2017, pp. 240–241.

[5] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H. J. Yoo, "UNPU: A 50.6TOPS/W unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision," in *2018 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2018, pp. 218–220.

[6] K. Ueyoshi, K. Ando, K. Hirose, S. Takamaeda-Yamazaki, J. Kadomoto, T. Miyata, M. Hamada, T. Kuroda, and M. Motomura, "QUEST: A 7.49TOPS multi-purpose log-quantized DNN inference engine stacked on 96MB 3D SRAM using inductive-coupling technology in 40nm CMOS," in *2018 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2018, pp. 216–218.

[7] B. Moons, B. D. Brabandere, L. V. Gool, and M. Verhelst, "Energy-efficient convnets through approximate computing," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016.

[8] L. Mei, M. Dandekar, D. Rodopoulos, J. Constantin, P. Debacker, R. Lauwereins, and M. Verhelst, "Sub-word parallel precision-scalable MAC engines for efficient embedded DNN inference," in *1st IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, March 2019.

[9] V. Camus, M. Cacciotti, J. Schlachter, and C. Enz, "Design of approximate circuits by fabrication of false timing paths: The carry cut-back adder," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, June 2018, pp. 746–757.

[10] H. Sharma, J. Park, N. Suda, L. Lai, B. Chau, J. K. Kim, V. Chandra, and H. Esmaeilzadeh, "Bit fusion: Bit-level dynamically composable architecture for accelerating deep neural networks," in *45th IEEE International Symposium on Computer Architecture (ISCA)*, June 2018, pp. 764–775.