14.5 ENVISION: A 0.26-to-10TOPS/W Subword-Parallel Dynamic-Voltage-Accuracy-Frequency-Scalable Convolutional Neural Network Processor in 28nm FDSOI

Bert Moons, Roel Uytterhoeven, Wim Dehaene, Marian Verhelst

KU Leuven, Leuven, Belgium

ConvNets, or Convolutional Neural Networks (CNN), are state-of-the-art classification algorithms, achieving near-human performance in visual recognition [1]. New trends such as augmented reality demand always-on visual processing in wearable devices. Yet, advanced ConvNets achieving high recognition rates are too expensive in terms of energy as they require substantial data movement and billions of convolution computations. Today, state-of-the-art mobile GPU's and ConvNet accelerator ASICs [2][3] only demonstrate energy-efficiencies of 10's to several 100's GOPS/W, which is one order of magnitude below requirements for always-on applications. This paper introduces the concept of hierarchical recognition processing, combined with the Envision platform: an energy-scalable ConvNet processor achieving efficiencies up to 10TOPS/W, while maintaining recognition rate and throughput. Envision hereby enables always-on visual recognition in wearable devices.

Figure 14.5.1 demonstrates the concept of hierarchical recognition. Here, a hierarchy of increasingly complex individually trained ConvNets, with different topologies, different network sizes and increasing computational precision requirements, is used in the context of person identification. This enables constant scanning for faces at very low average energy cost, yet rapidly scales up to more complex networks detecting a specific face such as a device's owner, all the way up to full VGG-16-based 5760-face recognition. The opportunities afforded by such a hierarchical approach span far beyond face recognition alone, but can only be exploited by digital systems demonstrating wide-range energy scalability across computational precision. State-of-the-art ASICs in references [3] and [4] only show 1.5× and 8.2× energy-efficiency scalability, respectively. Envision improves upon this by introducing subword-parallel Dynamic-Voltage-Accuracy-Frequency Scaling (DVAFS), a circuit-level technique enabling 40× energy-precision scalability at constant throughput. Figure 14.5.2 illustrates the basic principle of DVAFS and compares it to Dynamic-Accuracy Scaling (DAS) and Dynamic-Voltage-Accuracy Scaling (DVAS) [4]. In DAS, switching activity and hence energy consumption is reduced for low precision computations by rounding and masking a configurable number of LSB's at the inputs of multiply-accumulate (MAC) units. DVAS exploits shorter critical paths in DAS's reduced-precision modes by combining it with voltage scaling for increased energy scalability. This paper proposes subword-parallel DVAFS, which further improves upon DVAS, by reusing inactive arithmetic cells at reduced precision. These can be reconfigured to compute 2×1-8b or 4×1-4b (N×1-16b/N, with N the level of subword-parallelism), rather than 1×1-16b words per cycle, when operating at less than 8b precision. At constant data throughput, this permits lowering the processor's frequency and voltage significantly below DVAS values. As a result, DVAFS is a dynamic precision technique which simultaneously lowers all run-time adaptable parameters influencing power consumption: activity α, frequency f and voltage V. Moreover, in contrast to DAS and DVAS, which only save energy in precision-scaled arithmetic blocks, DVAFS allows lowering f and V of the full system, including control units and memory, hereby shrinking non-compute energy overheads drastically at low precision.

Energy-efficiency is further improved by modulating the body bias (BB) in an FDSOI technology. This permits tuning of the dynamic vs. leakage power balance while considering the computational precision. At high precision, reducing V_t allows a scaling down of the supply voltage to reduce dynamic consumption while maintaining speed, at a limited leakage energy cost and an overall efficiency increase. At low precision, and reduced switching activity, V_t and the supply voltage are increased to lower the leakage overhead at constant speed. This increases dynamic energy, but reduces the overall energy consumption.

Figure 14.5.3 shows the top-level architecture of Envision. This chip is a multi-power and multi-body-bias domain, sparsity-guarded ConvNet processor exploiting DVAFS. It is fully C-programmable, allowing deployment of a wide range of ConvNet topologies, and has a 16b SIMD RISC instruction set extended with custom instructions, similar to [4]. The processor is equipped with 2D- (for convolutions) and 1D-SIMD arrays (for ReLU, max-pooling) and a scalar unit. An on-chip memory (DM) consists of 64×2kB single-port SRAM macros, subdivided into 4 blocks of 16 parallel banks, storing a maximum of 65536×N words. 3 blocks can be read or written in parallel: 2 blocks by the processor, another by the Huffman DMA, used for compressing IO bandwidth up to 5.8×. The system is divided into three power- and body-bias domains to enable granular dynamic voltage scaling.

Figure 14.5.4 shows how the 6-stage pipelined processor executes convolutions in its 16×16 2D-SIMD MAC array. Each MAC is a single cycle N-subword-parallel multiplier, followed by a N×48b/N reconfigurable accumulation adder and register. As such, the 16×16 array can generate N×256 intermediate outputs per cycle while consuming only N×16 filter weights and N×16 features in a first convolution cycle. In subsequent cycles, a 256b FIFO further reduces memory bandwidth by reusing and shifting features along the x-axis, requiring only a single new feature fetch per cycle. As all intermediate output values are stored in accumulation registers, there is no data-transfer between MACs and no frequent write-back to SRAM. Sparsity is exploited by guarding both memory fetches and MAC operations [4], using flags stored in a GRD memory. This leads to an additional 1.6× system-wide gain in energy consumption compared to DVAFS alone for typical ConvNets (30-60% zeroes).

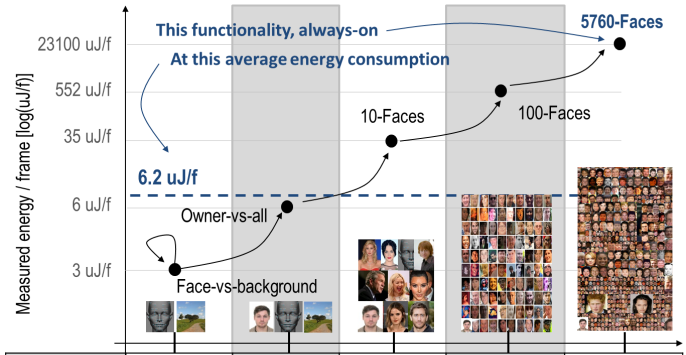
Envision was implemented in a 28nm FDSOI technology on 1.87mm^2 and runs at 200MHz at 1V and room temperature. Fig. 14.5.5 shows measurement results highlighting its wide-range precision-energy scalability, with nominal and optimal body-biasing. All modes run the same 5×5 ConvNet-layer, with a typical MAC efficiency of 73%, or $0.73×f×N×256×2$ effective operations per second. When scaling down from 16- to 4×4b sparse computations at 76GOPS, power goes from 290mW down to 7.6mW, as supply voltage and body bias are modulated between 0.65-1.1V and ±0.2-1.2V. Measurements for the convolutional layers in hierarchical face recognition are listed in Fig. 14.5.1, demonstrating 6.2μJ/f at average 6.5mW instead of 23100μJ/f at 77mW. This illustrates the feasibility of always-on recognition through hierarchical processing on energy-scalable Envision.

Figure 14.5.6 shows a comparison with recent ConvNet ASICs. Envision scales efficiency on the AlexNet convolutional layers between 0.8-3.8TOPS/W, compared to 0.16TOPS/W [3] and 0.56-1.4TOPS/W [4]. Efficiency is 2TOPS/W average for VGG-16 and up to 10 TOPS/W peak. This further illustrates Envision's ability to minimize energy-consumption for any ConvNet, demonstrating an energy-scalability of up to 40× at nominal throughput in function of precision and sparsity, hereby enabling always-on hierarchical recognition.

Figure 14.5.7 shows a die photo of Envision, illustrating the physical placement of its 3 power domains in a 1.29×1.45mm^2 active area.

Acknowledgements:
This work is partially funded by FWO and Intel Corporation. We thank Synopsys for providing their ASIP Designer tool suite and STMicroelectronics for silicon donation. Special thanks to CEA-LETI and CMP for back-end support.

References:
[1] Y. LeCun, et al., "Deep Learning," *Nature*, vol. 512, no. 7553, pp 436-444, 2015.
[2] L. Cavigelli, et al., "Origami: A Convolutional Network Accelerator," *IEEE Great Lakes Symp. on VLSI,* 2015.
[3] Y.H. Chen, et al., "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *ISSCC,* pp. 262-263, 2016.
[4] B. Moons, et al., "A 0.3-2.6 TOPS/W Precision-Scalable Processor for Real-Time Large-Scale ConvNets," *IEEE Symp. VLSI Circuits*, 2016.



Time-on	98%	1%	0.9%	0.09%	0.01%
Complexity	6M MACs	12M MACs	30M MACs	500M MACs	15.4G MACs
Sparsity	5-44%	8-45%	10-33%	8-47%	5-82%
Accuracy	2-4b (N=4)	3-4b (N=4)	4-6b (N=2)	4-6b (N=2)	4-6b (N=2)
Conv. model size	22kB (on-chip)	42kB (on-chip)	112kB	742kB	15MB
Recognition rate	94.3%	95.9%	95.4%	94%	90% - 95%
Chip efficiency	4.2 TOPS/W	4 TOPS/W	1.8 TOPS/W	1.8 TOPS/W	1.3 TOPS/W
Frequency	25 MHz	25 MHz	50 MHz	50 MHz	200 MHz
Avg. Power	6.4 mW	6.6 mW	15 mW	14.4 mW	77 mW

Figure 14.5.1: Hierarchical face recognition.

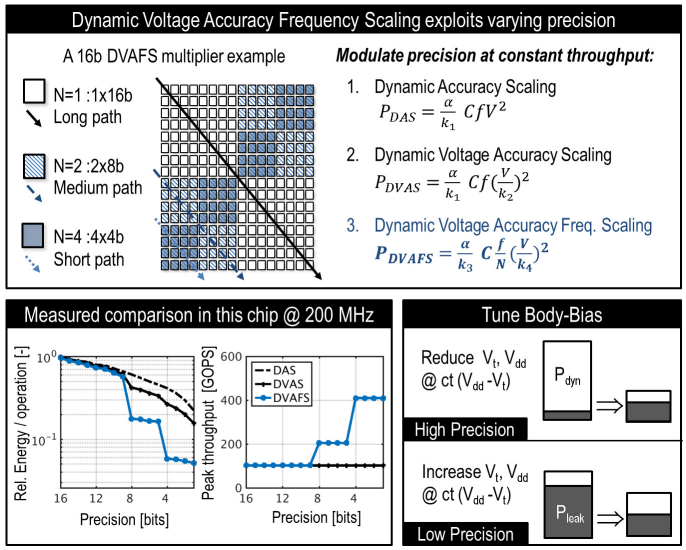


Figure 14.5.2: DVAFS and body bias tuning.

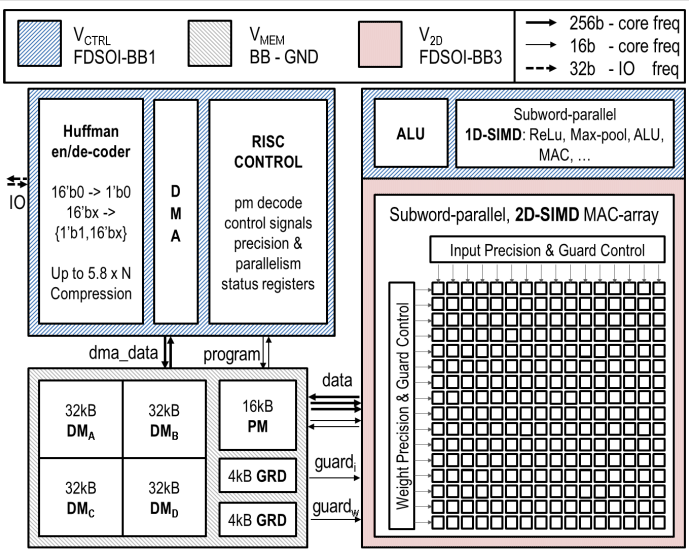


Figure 14.5.3: Top-level architecture of Envision.

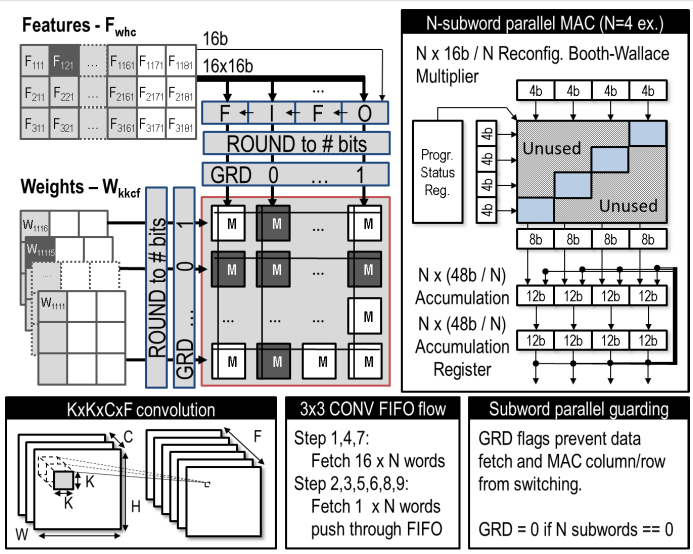


Figure 14.5.4: Parallel, rounded and guarded data flow.

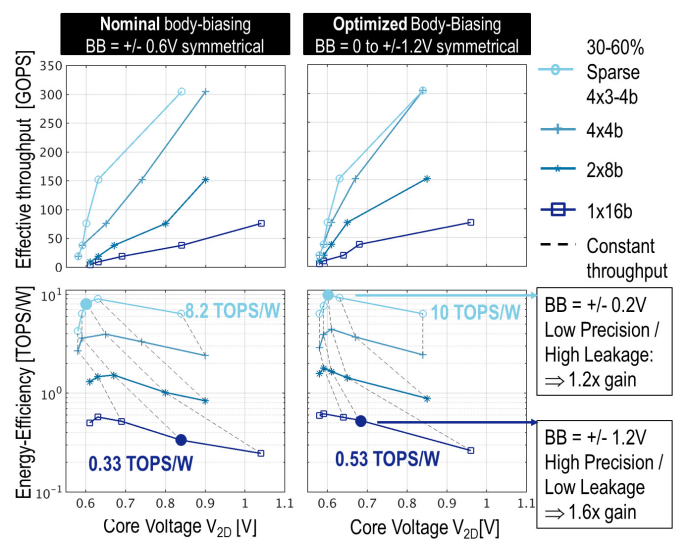


Figure 14.5.5: Measured efficiency up to 10 TOPS/W.

	[2] GLSVLSI '15	[3] ISSCC '16	[4] VLSI '16	**This work N = 1, 2 or 4**
Technology	65nm CMOS	65nm LP CMOS	40nm LP CMOS	**28nm UTBB FD-SOI**
Nominal Frequency [MHz]	500	200	200	**200**
Supply @ f_{nom} [V]	1.2	1	1.1	**1**
Peak performance [GOPS]	Fixed 196	Fixed 67	Fixed 102	**Dynamic N x 102**
Active Area [mm²]	1.31	12.25	2.4	**1.87**
# of MACs	-	168	256	**Dynamic N x 256**
Gate Count [NAND-2]	0.9M	1.852M	1.6M	**1.95M**
On-Chip SRAM [kB]	43	184.5	144	**144**
# layers, # filters [-]	All	All	All	**All**
Filter sizes [-]	<7x7	1-1024	All	**All**
Precision [bits]	Fixed 12	Fixed 16	Dynamic 1-16	**Dynamic N x 1-16 / N**
AlexNet Conv-layers [mW]	-	278 @ 34.7fps	55-95, 76 avg. @ 47fps	**20-62, 44 avg. @ 47fps**
VGG Conv-layers [mW]	-	-	-	**19-35, 26 avg. @ 1.67fps**
Dynamic power range @ $GOPS_{nom}$[mW]	510 (1x) @ 145 GOPS	235-332 (1.5x) @ 51-56 GOPS	35-288 (8.2x) @ 80 GOPS	**7.5-300 (40x) @ 76 GOPS**
Min. efficiency [TOPS/W]	0.44	0.15	0.27	**0.26**
Max. efficiency [TOPS/W]	0.8	0.35	2.6	**10**

Figure 14.5.6: Embedded ConvNet comparison.

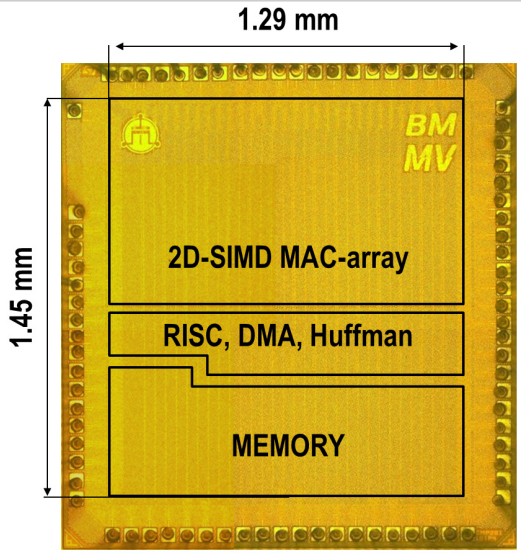


Figure 14.5.7: Die micrograph of Envision.

978-1-5090-3758-2/17/$31.00 ©2017 IEEE