



HAL
open science

A 35.6TOPS/W/mm² 3-Stage Pipelined Computational SRAM with Adjustable Form Factor for Highly Data-Centric Applications

J.-P Noel, M. Pezzin, R. Gauchi, J.-F Christmann, M. Kooli, Henri-Pierre Charles, L. Ciampolini, M. Diallo, F. Lepin, B. Blampey, et al.

► To cite this version:

J.-P Noel, M. Pezzin, R. Gauchi, J.-F Christmann, M. Kooli, et al.. A 35.6TOPS/W/mm² 3-Stage Pipelined Computational SRAM with Adjustable Form Factor for Highly Data-Centric Applications. IEEE Solid-State Circuits Letters, 2020, 10.1109/LSSC.2020.3010377 . cea-02904882

HAL Id: cea-02904882

<https://cea.hal.science/cea-02904882>

Submitted on 22 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A 35.6TOPS/W/mm² 3-Stage Pipelined Computational SRAM with Adjustable Form Factor for Highly Data-Centric Applications

J.-P. Noel¹, M. Pezzin¹, R. Gauchi¹, J.-F. Christmann¹, M. Kooli¹, H.-P. Charles¹, L. Ciampolini¹, M. Diallo¹, F. Lepin¹, B. Blampey², P. Vivet¹, S. Mitra³ and B. Giraud¹
¹Univ. Grenoble Alpes, CEA, LIST, F-38000 Grenoble, France
²Univ. Grenoble Alpes, CEA, LETI, F-38000 Grenoble, France
³Stanford University, Stanford, USA
 jean-philippe.noel@cea.fr

Abstract—In the context of highly data-centric applications, close reconciliation of computation and storage should significantly reduce the energy-consuming process of data movement. This paper proposes a Computational SRAM (C-SRAM) combining In- and Near-Memory Computing (IMC/NMC) approaches to be used by a scalar processor as an energy-efficient vector processing unit. Parallel computing is thus performed on vectorized integer data on large words using usual logic and arithmetic operators. Furthermore, multiple rows can be advantageously activated simultaneously to increase this parallelism. The proposed C-SRAM is designed with a two-port pushed-rule foundry bitcell, available in most existing design platforms, and an adjustable form factor to facilitate physical implementation in a SoC. The 4kB C-SRAM testchip of 128-bit words manufactured in 22nm FD-SOI process technology displays a sub-array efficiency of 72% as well as an additional computing area of less than 5%. The measurements averaged on 10 dies at 0.85V and 1GHz demonstrate an energy efficiency per unit area of 35.6 and 1.48TOPS/W/mm² for 8-bit additions and multiplications with 3ns and 24ns computing latency, respectively. Compared to a 128-bit SIMD processor architecture, up to 2x energy reduction and 1.8x speed-up gains are achievable for a representative set of highly data-centric application kernels.

I. INTRODUCTION

In- and Near-Memory Computing (IMC/NMC, depending on pre/post sense-amplifier computation) are considered today as promising solutions to overcome the energy wall problem raised with highly data-centric applications [1-2]. This problem corresponds to the huge amount of energy to move the data between memory and processing element compared to the small amount of energy required for the computation itself. Moreover, the technology scaling only exacerbates this phenomenon because although it reduces the computing energy, it increases the impact of moving data (by the increase of parasitic elements) that is predominant [3]. Recent works demonstrate the interest to use SRAM-based IMC architectures enabling multi-row selection during the read operation to perform *in-situ* logic operations (NOR, AND...), also called scouting logic [3-5]. This leads to significant energy reductions for representative sets of edge-AI and security-oriented kernels. Nevertheless, most of them require custom bitcells designed at logic-rules, at the cost of a drastic loss in memory density. Furthermore, to perform

bitwise IMC operations, the operands must be physically aligned at array level. This forbids to interleave the words on a single row, fixing the macro form factor to MUX-1.

In this paper, we propose a 128-bit Computational SRAM (C-SRAM) architecture with an adjustable form factor to facilitate physical implementation and compatible with two-port (8T) pushed-rule (SRAM-rules) foundry bitcells to keep a high memory density. Furthermore, we have paid particular attention to minimize the additional computing area and the performance penalty, while limiting the computing latency with a 3-stage pipeline. The main contributions of the paper are:

- To quantify the energy and power contributions (computing, fetch, store and NOP) of IMC and NMC operations with a 3-stage pipeline.
- To demonstrate the feasibility of designing high-density C-SRAM with adjustable form factor.

The remainder of the paper is organized as follow: Section II details the proposed 128-bit C-SRAM architecture as well as IMC/NMC operations and the associated pipeline flow. Section III explains the circuit design choices enabling the form factor adjustment to facilitate physical implementation. Then, the measurement results are depicted and compared with the prior art in Section IV. Application benchmarks versus 128-bit SIMD processor are also presented. Finally, Section V summarizes the paper.

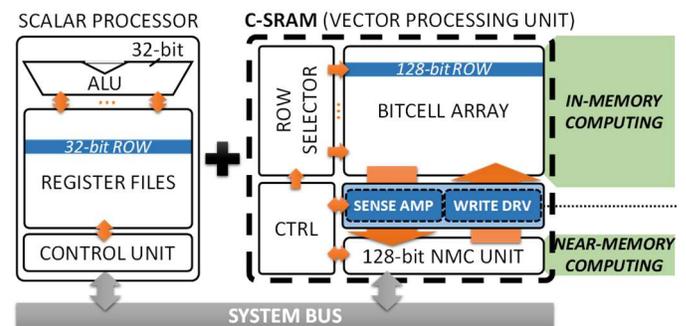


Fig. 1. Integration scheme of the proposed C-SRAM used as an energy-efficient vector processing unit.

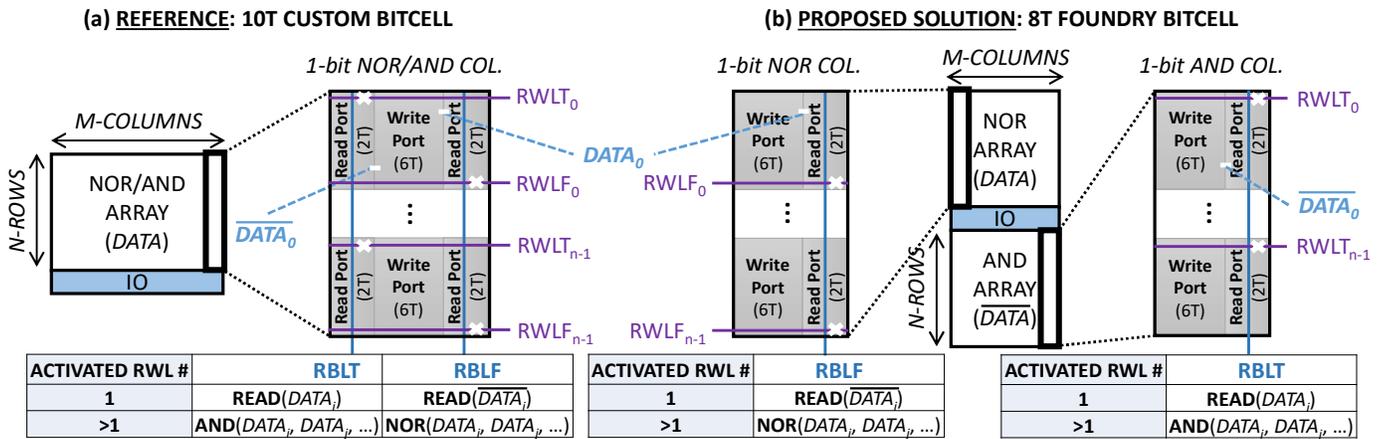


Fig. 2. (a) Reference (10T bitcell) and (b) proposed IMC solution based on 8T foundry bitcells to implement dual arrays enabling both NOR and AND operations.

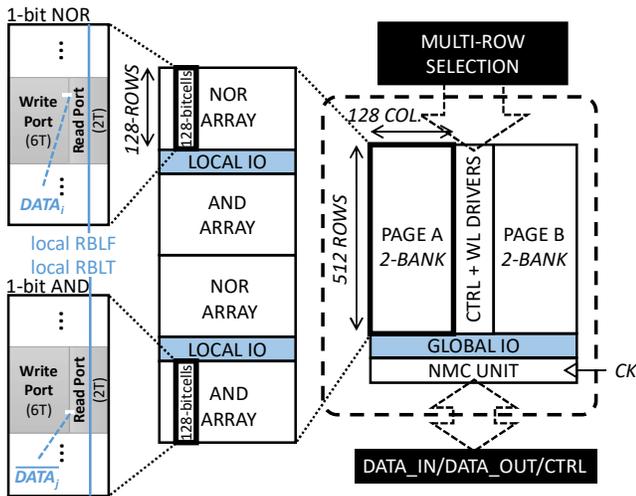


Fig. 3. Proposed C-SRAM architecture based on two adjacent memory pages.

II. PROPOSED C-SRAM ARCHITECTURE

The proposed C-SRAM architecture can be used to design an energy efficient vector processing unit performing heavily parallel operations which can be finely interleaved with sequential operations of a scalar processor (Fig. 1). Fig. 2(a) shows the baseline IMC approach based on a three-port (10T) custom bitcell [5], which is not supported by most of the foundries in advanced technology nodes. The proposed IMC approach (Fig. 2(b)) is based on dual arrays of two-port (8T) foundry bitcells per bank containing complementary data to enable simultaneously NOR and AND operations when multi-row selection is activated. Fig. 3 shows the proposed C-SRAM

architecture including 2 memory banks in MUX-1 configuration (no word interleaving). Each bitcell array (NOR & AND) is connected to a Local I/O (LIO), while a Global I/O (GIO) vertically interfaces the 2 banks of the page with a vectorized NMC unit. Based on this architecture, a representative set of logic and arithmetic operations is supported (Fig. 4(a)). To ensure conflict-free memory access, while performing energy-efficient operations, the six types of C-SRAM instruction are executed in a 3-stage pipeline (Fig. 4(b)). Depending on the instruction sequence (Fig. 4(c)), minimum and maximum power consumption occur when the 3 pipeline stages are either idle (NOP) or simultaneously operating (IMC+NMC+WR), respectively.

III. ADJUSTABLE FORM FACTOR DESIGN & IMPLEMENTATION

A 256x128b (4kB) C-SRAM testchip has been manufactured in 22nm FD-SOI process technology, featuring 1 page of 2 banks organized in dual arrays of 128-rows of 128-columns. In order to demonstrate the feasibility of designing a C-SRAM with an adjustable form factor, an extra memory page including only interconnecting layers taking the parasitic elements into account has been implemented. To ensure IMC operations (NOR/AND) between the two adjacent pages, an original horizontal BL metallization scheme is used to connect together the vertical Global BLs of each bank to a Vertical I/O (VIO) (Fig. 5). Multi-operand IMC is then enabled by bitwise row selection inputs providing different test scenarios (each bit activates a WL). The proposed metallization scheme only requires 2 additional metal layers compared to a conventional SRAM in the same process technology.

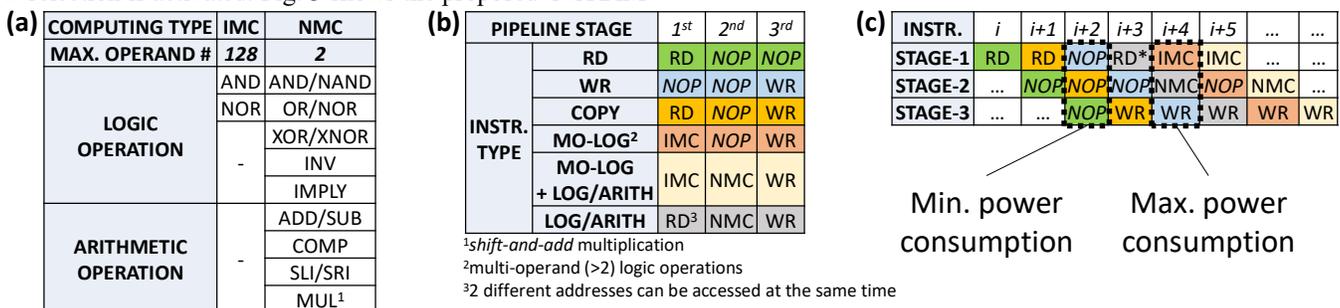


Fig. 4. IMC & NMC operations: (a) supported functions, (b) instruction types and (c) instruction flow example.

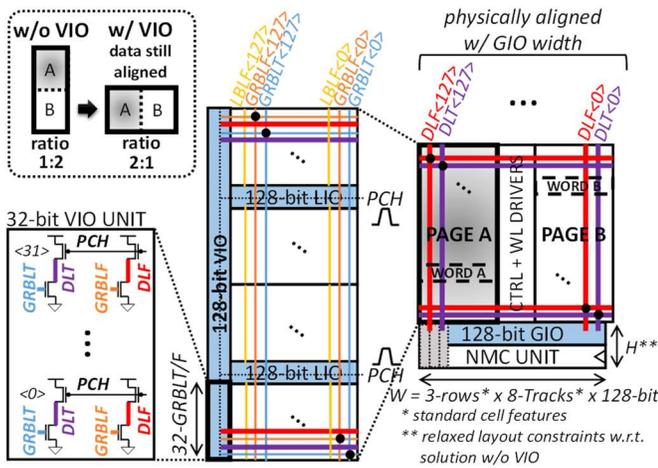


Fig. 5. Horizontal BL metallization scheme enabling adjustable form factor for inter-page IMC operands (e.g. ‘word_A’ AND ‘word_B’).

IV. MEASUREMENT RESULTS & APPLICATIONS

All reported data are measurements averaged on 10 dies. Fig. 6 shows that the computing part performed in the second stage of the pipeline (operation execution) represents more than 40% of the total instruction energy. The rest is spent by fetching the operands (~30%) and storing the results (~30%). Note that the difference between logic and arithmetic operations (except for multiplication) is negligible (<4%). These results highlight the advantages of performing NMC operations when the data to be processed is already in memory. Fig. 7 represents the IMC operation (NOR/AND) energy according to the number of activated RWL (corresponding to operands). The difference between 2 and 128 activated RWL is below 20% because most of the energy is spent in the IOs (local, global and vertical) and the controller. Fig. 8(a) shows that the instruction energy ranges from 118 to 211 fJ/bit (1.79x) at 0.85V according to the instruction type. Fig. 8(b) lists the power consumption for different pipeline filling scenarios, varying from 11 to 27.1mW (2.46x) at 0.85V and 1GHz. Using a 3-stage pipeline allows to limit the variation in energy and power consumption, which will make easier the power management at the system level.

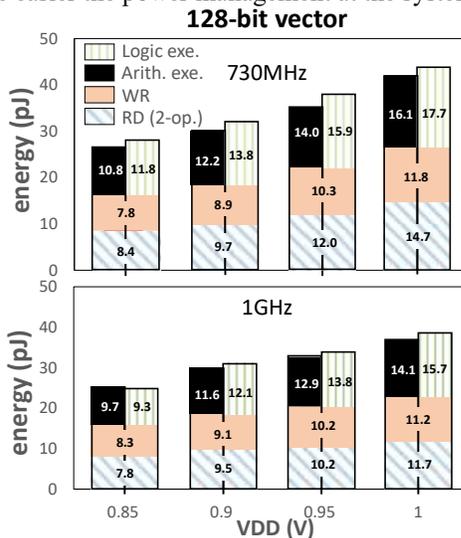


Fig. 6. Measured logic and arithmetic instruction energy vs. VDD.

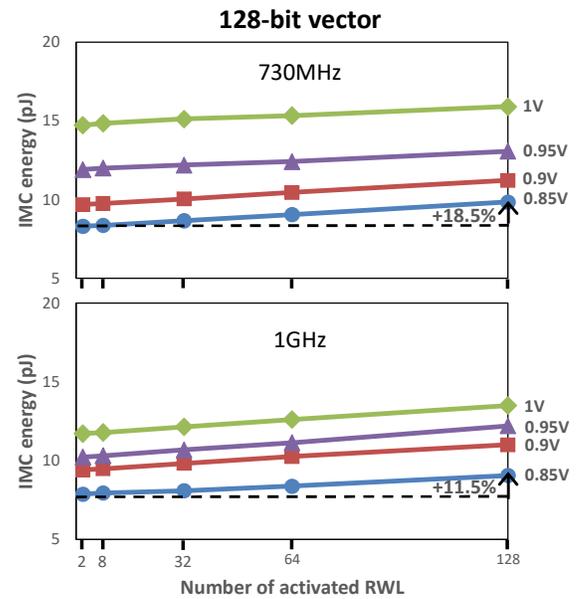


Fig. 7. Measured IMC energy vs. number of activated RWL.

To evaluate the benefits at architecture level, the C-SRAM is compared to a 128-bit SIMD processor architecture (using Intel SSE instructions) for two representative application kernels: Shift-OR (database-oriented) and Boolean Matrix Multiplication (AI-oriented) using a simulation platform. Fig. 9 describes this SystemC/TLM-based platform power-annotated with silicon measurements of the C-SRAM and a baseline 128-bit SIMD processor architecture [7]. Fig. 10 shows gains up to 1.8x in speed-up, 2x in energy reduction and 3.7x in energy-delay product (EDP) at 0.85V and 1GHz. Table I compares the C-SRAM to previous works and shows significant improvements in terms of memory density (up to 1.87Mb/mm²), form factor flexibility, computing latency (down to 3ns) and energy efficiency per unit area (up to 30.5TOPS/W/mm²) for the highest frequency (1GHz). For the sake of fairness, energy-related performance is compared at same VDD (0.9V). Fig. 11 shows the die micrograph and the testchip summary at nominal VDD (0.85V). In this condition, the energy efficiency per unit area is increased by 16% to achieve 35.6 and 1.48TOPS/W/mm² for 8-bit additions and multiplications, respectively.

INSTR. TYPE	PIPELINE STAGE			Energy (fJ/bit) @0.85V	
	1	2	3	Min.	Max.
RD	RD	NOP	NOP	118	
WR	NOP	NOP	WR	122	
COPY	RD	NOP	WR	154	
MO-LOG*	IMC	NOP	WR	155	164
MO-LOG + LOG/ARITH	IMC	NMC	WR	199	211
LOG/ARITH	RD	NMC	WR	199	202

PIPELINE FILLING SCENARIO			Power (mW) @0.85V/1GHz
STAGE-1	STAGE-2	STAGE-3	
NOP	NOP	NOP	11
RD (2-OPERANDS)	NMC	WR	25.5
IMC (2-OPERANDS)			25.5
IMC (128-OPERANDS)			27.1

*multi-operand (>2) logic operations

Fig. 8. Measured (a) instruction energy and (b) pipeline power consumption.

TABLE I. COMPARISON WITH PREVIOUS WORKS @0.9V

	This Work	JSSC 2020 [4]	JSSC 2018 [5]	VLSI 2019 [6]
Process technology node	22nm	28nm	40nm	65nm
SRAM bitcell	foundry (SRAM-rules) 8T	custom (logic-rules) 8T	custom (logic-rules) 10T	custom (logic-rules) 6T
Memory capacity	4kB	4kB (instance) 128kB (full chip)	8kB	8kB
Memory density	1.87Mb/mm ²	0.816Mb/mm ²	0.3Mb/mm ²	0.01Mb/mm ²
Adjustable form factor	yes	no	no	no
Computing type	IMC+NMC (bit-parallel)	IMC+NMC (bit-serial)	IMC+NMC (bit-parallel)	IMC+NMC (bit-parallel)
Supported IMC operation	AND/OR (multi-rowselection)	AND/OR (multi-rowselection)	AND/OR (2-rows selection)	XNOR
Supported NMC operation	Logic/Add/Sub/Comp/Shift/Mul	Logic/Add/Sub/Comp/Mul/Div/FP	Logic/Shifter/Rotator/S-BOX	MAC
Targeted usage	vector processing unit	vector processing unit	accelerator (cryptography only)	accelerator (RNN only)
Operating frequency	1GHz @0.9V	375MHz @0.9V	90MHz @0.9V	75MHz @0.9V
Peak performance (GOPS)	16 ^a (8-bit ADD) 0.67 ^a (8-bit MUL)	3 ^b (8-bit ADD) 0.23 ^b (8-bit MUL)	96 ^b (8-bit ADD) 7.5 ^b (8-bit MUL)	614
Computing latency @Op. freq. (ns)	3 ^a (8-bit ADD) 24 ^a (8-bit MUL)	21 (8-bit ADD) 271 (8-bit MUL)	not reported	not reported
Max. energy per computed bit (fJ/bit)	248 @0.9V ^a	1536 @0.9V ^b	48 @0.9V ^b	152 @0.9V ^c
Peak energy efficiency per unit area (TOPS/W/mm ²)	30.5 ^a (8-bit ADD) 1.27 ^a (8-bit MUL)	0.08 (8-bit ADD) ^b 0.009 (8-bit MUL) ^b	2.6 (8-bit ADD) ^b 0.28 (8-bit MUL) ^b	not reported
				2.7 @0.9V ^d

^a constant whatever computing load ^b peak performance when the pipeline is full ^c assuming 128-bit words ^d assuming 256-bit words

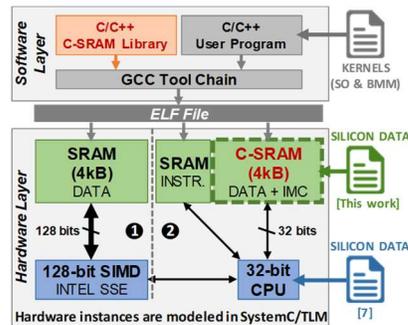


Fig. 9. Benchmark platform using SystemC/TLM calibrated with silicon data.

V. CONCLUSION

This paper presents a 4kB C-SRAM circuit based on two-port pushed-rule foundry bitcells manufactured in 22nm FD-SOI process technology. The testchip achieves the highest memory density (1.87Mb/mm²) compared to previous works, while facilitating physical implementation at SoC level thanks to a novel design technique enabling the form factor adjustment. Based on a 3-stage pipeline, access to memory as well as execution of operations (IMC/NMC) can operate up to 1 GHz with a latency of 3ns (except for 8-bit multiplication which has a 24ns latency). This small pipeline depth limits the energy (118 to 211 fJ/bit) and power consumption (11 to 27.1mW) variations as well as the additional area of the unit managing operations (<5%). Thanks to the combined benefits of all these circuit design choices, the proposed C-SRAM solution achieves an energy efficiency per unit area of 35.6 and 1.48TOPS/W/mm² at 0.85V for 8-bit additions and multiplications, respectively. At architecture level, compared to a 128-bit SIMD processor dealing with a 4kB data set, up to 3.7x EDP is obtained with a representative set of edge-AI and database application kernels.

ACKNOWLEDGMENTS

This work was supported by French ANR via Carnot funding.

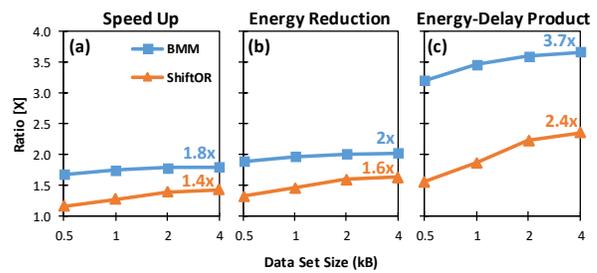


Fig. 10. Simulated (a) speed-up, (b) energy reduction and (c) EDP ratio of 128-bit C-SRAM vs. 128-bit SIMD processor (baseline).

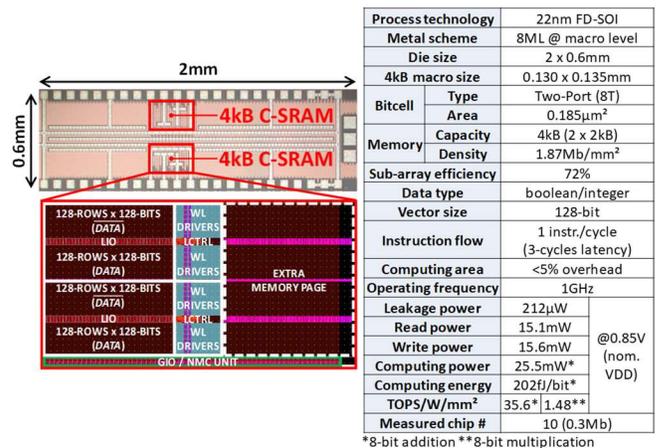


Fig. 11. Die micrograph and chip summary @0.85V

REFERENCES

- [1] N. Verma *et al.*, SSCM, 2019, pp. 43-55.
- [2] H. Valavi *et al.*, JSSC, 2019, pp. 1789-1799.
- [3] M. Horowitz, ISSCC, 2014, pp. 10-14.
- [4] J. Wang *et al.*, JSSC, 2020, pp. 76-86.
- [5] Y. Zhang *et al.*, JSSC, 2018, pp. 217-230.
- [6] R. Guo *et al.*, VLSI, 2019, pp. 995-1005.
- [7] J.-F. Christmann *et al.*, JLPEA, 2019