Impact of Analog Non-Idealities on the Design Space of 6T-SRAM Current-Domain Dot-Product Operators for In-Memory Computing

Adrian Kneip, *Student Member, IEEE*, and David Bol, *Senior Member, IEEE*

***Abstract*— In-memory computing provides unprecedented power and area efficiency for the execution of convolutional neural networks by using memory bitcells to perform dot-product (DP) operations in the analog domain. Yet, these operators suffer from analog non-idealities (ANIs) that degrade the inference accuracy. This paper proposes design guidelines inferred from a holistic simulation-based analysis of the impact of ANIs on the accuracy-efficiency trade-off that affects current-domain DP operators based on conventional 6T-SRAM bitcell arrays. We define a custom SNR metric aware of the DP operand distribution to quantify decision errors associated with various ANIs, over ranges of input/output resolution and hardware design parameters. We find out that non-linearity and local mismatch are the dominant ANIs limiting the design space, while IR drops turn out to be critical only when targeting high parallelism. We then quantify the accuracy-efficiency trade-off related to these dominant ANIs across the design space and propose optimal design choices. We notably identify that using larger operators can either improve or worsen the SNR depending on the target output resolution. Furthermore, we show that hardware calibration techniques which mitigate mismatch help to recover a fraction of the lost SNR, with greater effectiveness when scaling down the supply voltage.**

***Index Terms*— 6T-SRAM, analog non-idealities, design space analysis, hardware calibration, in-memory computing, neural networks, quantization.**

I. Introduction

GROWING interest in processing machine learning (ML) tasks directly at the edge has raised unprecedented challenges in terms of computational efficiency amid concerns of bandwidth, energy and privacy [1]. In order to address these challenges, new computing paradigms have emerged, favouring local computations and data reuse [2], such as neuromorphic systems [3]. Computing in-memory (CIM) has recently delivered tremendous efficiency for the execution of convolutional neural networks (CNNs) by mapping their massively parallel dot-product (DP) operations directly inside memory arrays. The bitcell-based analog DP operators allow nowadays SRAM-based CIM accelerators (CIM-SRAMs) in standard CMOS processes to be $10-100\times$ more power and area efficient than state-of-the-art digital CNN accelerators [Fig. 1 (top)]. However, analog operators suffer from analog non-idealities (ANIs), exacerbated in dense memory environments. These ANIs create bit errors on the digitized DP result, which lead to a degradation of the inference accuracy. While ANI-aware training helps to recover part of this degradation [4], [5], its effectiveness is eventually limited by the hardware equivalent noise level [6]. The sensitivity to this noise is a function of the *design space* of the DP operators and the target input/output resolution. Nowadays, CIM-SRAMs strive for more resolution, required to obtain golden levels of inference accuracy on datasets of moderate complexity at the edge [5], [7], unreachable with efficient binary networks [8], as seen in Fig. 1 (bottom) for the CIFAR-10 dataset.

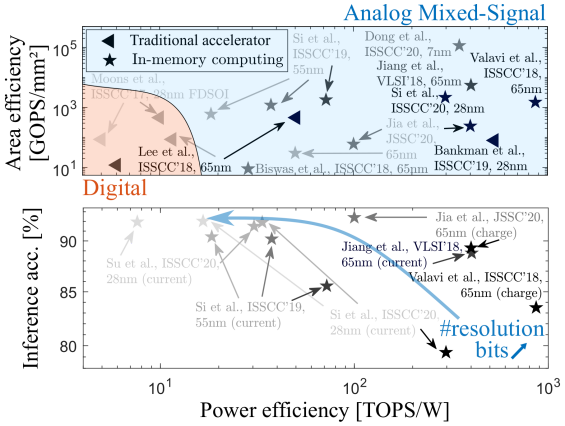

Fig. 1. (Top) Comparison of power/area efficiencies between state-of-the-art digital accelerators and analog mixed-signal CIM-SRAMs. (Bottom) Trade-off between power efficiency and inference accuracy (CIFAR-10 dataset), driven by the increase in CIM-SRAM output resolution (dark to light grey).

In this context, analyzing the impact of ANIs *at the hardware-level* is key to understand the physical impact on the accuracy-efficiency trade-off observed *at the system-level*. Yet, there have been few exhaustive analyses of ANIs over the broad design space of CIM-SRAMs in the literature. Kang *et al.* proposed an analytical model assessing the bit error probability due to non-linearity and hardware equivalent noise (mainly from local mismatch between transistors) on the output of their DIMA architecture [9]. They showed there exist fundamental limits to the accuracy level of mapped ML algorithms depending on the noise level. However, they rely on several analytical hypotheses and only detail the effects of changing the bitline voltage swing and operator

Manuscript received October 5, 2020; revised December 29, 2020 and February 1, 2021; accepted February 5, 2021. Date of publication February 24, 2021; date of current version April 27, 2021. This work was supported by the Fonds National de la Recherche Scientifique (FNRS), Belgium. This article was recommended by Associate Editor M.-F. Chang. *(Corresponding author: Adrian Kneip.)*

The authors are with the ICTEAM Institute, UCLouvain, 1348 Ottignies-Louvain-la-Neuve, Belgium (e-mail: adrian.kneip@uclouvain.be; david.bol@uclouvain.be).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2021.3058510.

Digital Object Identifier 10.1109/TCSI.2021.3058510

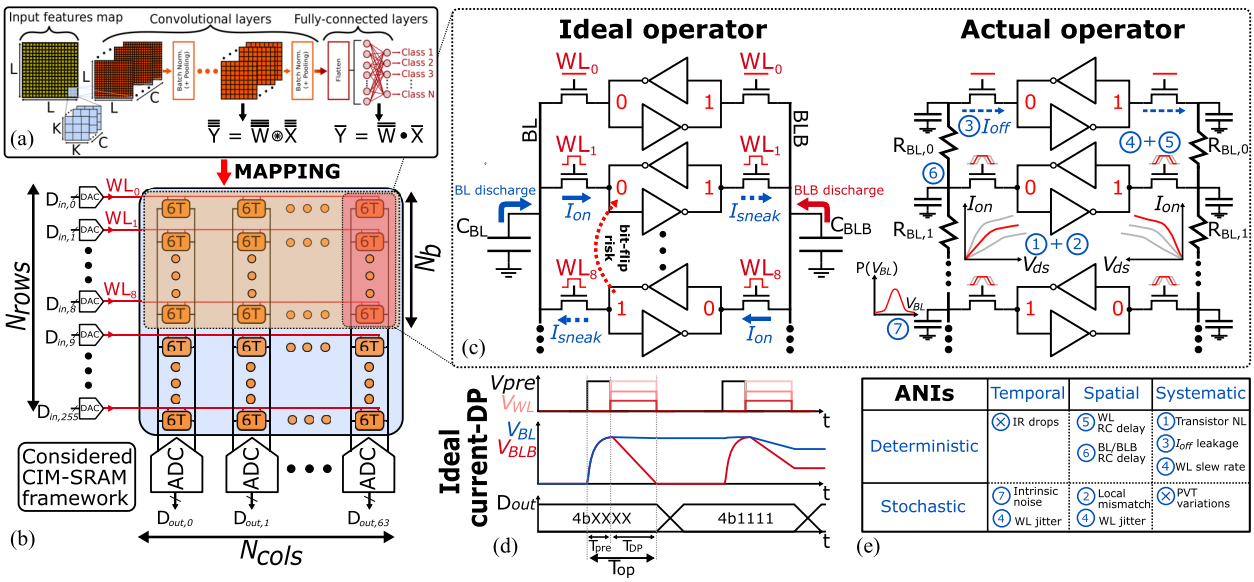

Fig. 2. General framework: (a) analysis setup, showing a simple CNN topology for image recognition, and (b) the considered CIM-SRAM architecture, able to map both convolutional and fully-connected operations on column-parallel DP current-domain operators. (c) Representation of the ideal and actual DP operators with (d) electrical waves depicting the ideal current-DP operation, neglecting sneak current and (e) classification of the considered ANIs.

size. Jaiswal *et al.* quantified mismatch and voltage drop on parasitic bitline resistances for 8T-based CIM-SRAMs with current-mode readout [10]. However, they do not provide a unified metric to compare these ANIs, nor do they consider output quantization. Also, they limit their analysis to small operator sizes ($\leq$ 16 parallel inputs). Ali *et al.* introduced an innovative multi-bit operator using parallel columns and discussed non-linearity, mismatch and process variations [11]. Yet, they do not present their sampling methodology and have only limited considerations for the effect of design parameters. Yin *et al.* provided hardware measurements highlighting the impact of ANIs with supply voltage for their XNOR-SRAM architecture [12]. They also simulated the relative impact of mismatch and bitline resistance with supply voltage, showing one or the other dominates as a function of it. Many works targeting emerging technologies addressed device modelling [6], [13] without considering quantization issues. Other works [14] developed system-level frameworks for the generic design of CIM architectures but focus on power/performance metrics. Finally, to the best of our knowledge, there has been no work addressing the effect of transistor V_t selection or technology scaling, nor comparing these many ANIs in a unified way.

In the present work, we propose a holistic analysis quantifying the impact of ANIs on the design space of in-memory current-domain DP operators based on 6T-SRAM bitcells (which we will now refer to as *6T current-DP operators*). We therefore divide the rest of this paper as follows. Section II presents the current-DP operation and the hypotheses underlying our analysis framework. We also introduce a custom *distribution-aware SNR* metric to objectively quantify bit errors at the output while accounting for the distribution of DP operands. Section III proposes mitigation schemes against the systematic ANIs that affect the DP operation. Section IV compares the impact of various spatial and temporal ANIs on the SNR metric and derives critical design space boundaries. Section V then analyzes the accuracy-efficiency trade-offs across the design space for the main sources of ANIs, deduced from Section IV. In Section VI, we discuss how hardware calibration techniques can push the design space boundaries from Section V. Eventually, we summarize and conclude.

II. Analysis Framework

Traditional CNNs as in Fig. 2 (a) consist of successive (2D) convolutional and fully-connected layers, separated by regularization and activation layers. Assuming fixed-point resolution, the neurons in these layers perform DP operations in parallel, yielding the *expected* outputs

$$\widehat{\mathrm{D}}_{out,n} = \rho\left(\widehat{\mathrm{DP}}_n\right) = \rho\left(\sum_{i=0}^{N_{in}-1} \mathrm{W}_{i,n}\ \mathrm{D}_{in,i}\right). \tag{1}$$

W and D_{in} respectively stand for the weights and digital inputs, $\widehat{\mathrm{DP}}$ is the expected DP output for a given (D_{in},W) combination and $\widehat{\mathrm{D}}_{out}$ is its quantized version in fixed-point representation, ρ is the output quantization function, and $N_{in/out}$ are respectively the number of inputs/outputs such that $n \in [0, N_{out}-1]$. *Please note that Appendix A summarizes the main recurrent symbols used in this work.* One can directly map these DP operations onto parallel DP operators inside an SRAM array, like the one in Fig. 2 (b). When mapping fully-connected layers, each column acts as a single neuron, with inputs fed horizontally and outputs retrieved vertically. For convolutional layers, filter weights are usually flattened in a single dimension to suit DP computations, each column mapping one output channel [15], [16]. In both cases, the computation relies upon the current-domain DP operators.

A. The Ideal 6T Current-DP Operator

Fig. 2 (c) and (d) show the ideal operator and depict the operation free of ANIs. Let N_b be the operator size and r_{in} the input resolution. First, N_b digital inputs are converted by

DACs into analog wordline (WL) voltages with $2^{r_{in}}$ levels. Meanwhile, bitlines (BLs/BLBs) are precharged to supply voltage level V_{DD}. The DP operation begins when N_{on} ($\leq N_b$) bitcells are asserted based on the WL values. For these N_{on} bitcells, the discharge current I_{on} depends on the WL level and the stored bitcell data, which acts as a differential $+1/-1$ weight W. The differential discharges of BLs/BLBs then build the analog DP voltages $V_{DP,n} = V_{BL,n} - V_{BLB,n}$ by the end of the DP duration, with $n \in [0, N_{cols} - 1]$. Eventually, column-pitched ADCs digitize these DP voltages with output resolution r_{out}, giving the *ideal* outputs

$$\begin{aligned} \mathrm{D}_{out,n} &= \rho_{ADC}\left(V_{DP,n}\right) \\ &= \rho_{ADC}\left(\frac{T_{DP}}{C_{BL}} \sum_{i=0}^{N_{on}-1} \mathrm{W}_{i,n}\, I_{on,i}(V_{WL,i})\right), \end{aligned} \quad (2)$$

with ρ_{ADC} the ADC quantization function, T_{DP} the duration of the DP operation, $C_{BL} = C_{BLB}$ the bitline capacitance and $I_{on,i}(V_{WL,i})$ the i-th pull-down current, whose value depends on its corresponding input-dependent WL voltage level. This expression matches the linear DP in Eq. (1), ignoring a multiplication factor and supposing the I_{on} values are linearly spaced following the DAC conversion. We investigate this point in Section III.B. Furthermore, we consider the following hypotheses during the CIM-SRAM study:

- We take $N_{rows} = 256$ and $N_{cols} = 64$ to allow the study of moderate-to-high operator sizes with reasonable computational cost. Hence, we can do up to 64 DP operations in parallel on the same 256-dimensional inputs.
- Given that recent designs in both current- [4], [11] and charge-domain [7] use multiple columns to express multi-bit weights followed by weighted analog or digital summation, we can study *binary +1/-1 weights in a single column* without major loss of generality.
- *Baseline conditions* use a 65nm LP CMOS technology with standard-V_t (SVT) transistors, supplied at their nominal 1.2V voltage. Input and output resolution are set to a default value of 1b.
- We use minimum-sized bitcells to keep maximum density, with a 1.5× upsized NMOS pull-down width to ensure read-stability.
- Total bitline capacitances $C_{BL(B)}$ account for layout-extracted parasitic values in metal-2 for 256 rows, which reach 50fF for the selected 65nm LP node.
- We assume ideal DAC and ADC operations as we focus here on the DP operator. By default, we assume that ADCs have an uniform quantization with a half-LSB negative voltage offset which removes the nominal deadband at 0V differential input (see Fig. 4). The DAC response will be detailed in Section III.B but works as a simple WL buffer for binary inputs.

B. Bit-1 Stability Conditions

CIM-SRAMs based on 6T bitcells suffer from sneak pull-up currents (I_{sneak} in Fig. 2) which can alter the DP result and jeopardize data stability, as first noted in [17]. Neglecting leakage, these currents arise when $V_{BL(B)} < V_{WL} - V_t$ and start to discharge the bit-1 node through the accessed bit-0 cells. This discharge can lead to undesired bit-flips if (i) the cell inverter PMOS cannot sustain the sneak current, and (ii) the DP duration is long enough for the bit-1 node capacitance to discharge below the inverter tipping point. Hence, at fixed supply voltage, the bit-1 stability is a function of V_{WL} and N_b, which control the driving strength of the access transistors and the maximum DP duration. Moreover, changing the DP duration also modifies the BL voltage swing. Because of high bitcell variability, we have to evaluate the statistical conditions yielding a 1-to-0 flip during the worst-case DP scenario: a single bitcell storing a bit-1 and $N_b - 1$ bitcells with a bit-0. Assuming that we strive for 99% of CIM-SRAMs without a single bit-flip, we should find the critical DP duration for which the bit-flip probability of an individual cell stays below $1 - 0.99^{1/(16\times256\times64)} = 3.84 \times 10^{-8}$ (0.03ppm), assuming 16 banks of size 256×64. We can accurately extract such rare events using high-sigma techniques, and resort here to the gradient-based importance sampling methodology from [18].

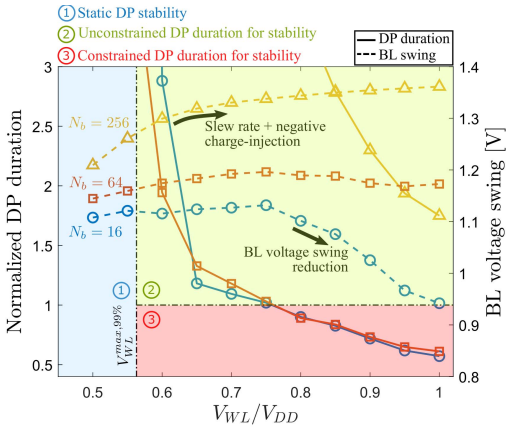

Fig. 3. Evolution of the normalized maximum DP duration for a 0.03ppm bit-flip with different WL voltage and operator sizes, at 1.2V in the 65nm LP node. The BL voltage swing under constrained conditions is also shown.

Fig. 3 features the ratio between the maximum DP duration that ensures a 0.03ppm bit-flip and the nominal DP duration to reach 95% of BL full-scale, at different values of WL voltage and 1.2V. It highlights three regions with different stability conditions: region ① ensures static stability of the data for any DP duration, region ② safely achieves the 0.03ppm target at the nominal DP duration and region ③ requires to shorten the maximum DP duration to ensure statistical retention. When the operator size increases, the nominal DP pulse width narrows and the bit-1 node has less time to discharge. Hence, the WL voltage at which we transition from region ② to ③ tends to rise, as seen for $N_b = 256$. Nonetheless, it becomes difficult to accurately monitor such narrow DP pulses in practice. We discuss this point in Section IV.E. Besides, we would expect a reduction of the maximum voltage swing when limiting the DP duration in region ③, as observed for $N_b = 16$. However, increasing the operator size appears to mitigate the swing reduction and even to increase it above full-scale. This happens due to an important one-sided negative charge-injection from the active WLs on the BL, worsening with higher WL voltage. Furthermore, the BL swing saturates as the DP duration becomes close to the considered 25ps slew rate. While one

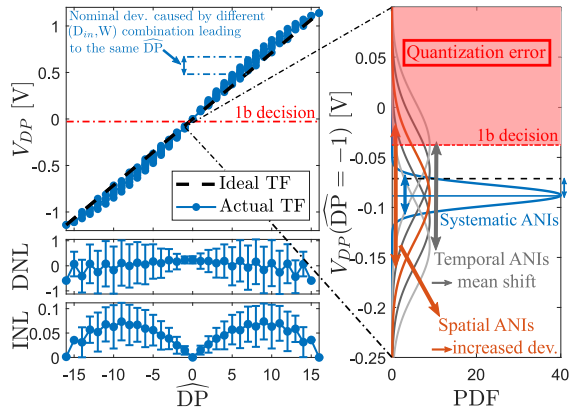

Fig. 4. Impact of ANIs on the DP transfer function (TF) for $N_b = 16$. (Left) Systematic ANIs alter the TF linearity, with the highest INL reached for intermediate $\widehat{\mathrm{DP}}$. The equivalence of (D_{in},W) combinations lead to nominal DNL (shown with 1-σ error bars). (Right) Spatial and temporal ANIs respectively increase the deviation and shift the mean of each $V_{DP}(\widehat{\mathrm{DP}})$ distribution, here for $\widehat{\mathrm{DP}} = -1$. These ANIs increase the likelihood of erroneous quantization near ADC thresholds, as seen when $r_{out} = 1$b.

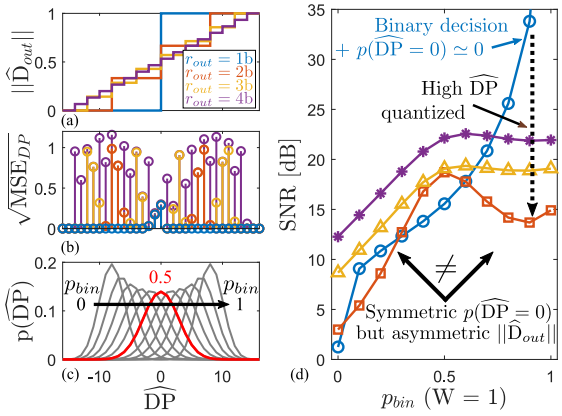

Fig. 5. Distribution-aware SNR components: (a) Ideal normalized ADC quantization function with increasing r_{out}. (b) Root MSE in baseline conditions (w/ mismatch). (c) $\widehat{\mathrm{DP}}$ distribution with gaussian D_{in} and binomial W sampling, increasing the binomial probability p_{bin}. (d) Obtained distribution-aware SNR metric with increasing p_{bin} and r_{out} in baseline conditions ($N_b = 16$), corresponding to the shown MSE data.

could take advantage of Fig. 3 for a dedicated design with fixed operator size, we opt for the conservative choice of fixing V_{WL} to the safe $V_{WL}^{max,99\%}$ static limit. Such choice minimizes the additional non-linear contribution from sneak currents and relaxes the design space exploration. However, note that this choice also reduces the maximum DP throughput and increases mismatch sensitivity compared to regions ② and ③.

C. ANIs Classification and Decision Errors

Despite stable bit-1 conditions, Eq. (2) cannot hold with a finite BL voltage swing and ANIs affecting the $V_{DP}(\widehat{\mathrm{DP}})$ transfer function. We sketch the impact of these ANIs on the DP operator in Fig. 2 (c). We did not draw IR drops for clarity. Note that charge-injection has eventually a low impact on the DP result thanks to the differential architecture. We classify these ANIs based on their *deterministic* or *stochastic* nature, as well as whether they induce the same constant error on all operators (systematic), whether they change the transfer function of a given operator over time (temporal) or due to bitcell variability and layout position dependence (spatial).

Fig. 4 (right) shows that systematic ANIs alter the linearity of the ideal voltage response. With the 25ps slew rate in baseline conditions, the observed non-linearity comes from the drain-to-source modulation of the I_{on} current by the BL voltage. Importantly, we also notice a nominal deviation of V_{DP} samples because different (D_{in},W) combinations can produce the same $\widehat{\mathrm{DP}}$ result (e.g. $\widehat{\mathrm{DP}} = 3 = 3 - 0 = 8 - 5$). These different combinations correspond to more or less bitcells passing, such that the differential error on the DP result changes, which spreads the nominal output. This notion of (D_{in},W) *equivalence* is a key concept in 6T-based architectures as it impacts both the differential and integrated non-linearity (DNL/INL) of the transfer function. Although $\widehat{\mathrm{DP}} = 0$ has the largest number of equivalent (D_{in},W), its DNL remains close to 0 because the corresponding differential voltage error is nominally very low. Moreover, BL clamping and a low number of equivalent (D_{in},W) combinations reduce errors at high $\widehat{\mathrm{DP}}$ values. Therefore, intermediate $\widehat{\mathrm{DP}}$ values show the highest DNL and INL. Besides, spatial and temporal ANIs further affect the nominal distribution of the DP results, as illustrated in Fig. 4 (right) for $\widehat{\mathrm{DP}} = -1$. Spatial ANIs, in particular local mismatch, increase the deviation of DP results due to position-dependent differences in individual I_{on} values. Temporal ANIs change the DP distribution in each operator over successive operations as noise varies over time. Altogether, spatio-temporal ANIs increase the likelihood of wrongly crossing one (or multiple) ADC decision threshold(s), as depicted for a 1b output. These incorrect ADC decisions are the eventual cause of bit errors on the digital outputs. In order to assess them quantitatively and to account for both the ADC resolution and operands distribution, we have to define an appropriate error metric and simulation framework.

D. Error Metric Definition

Let us consider the quantization of the previous DP results by the quantization steps in Fig. 5 (a). The resulting mean-square error (MSE) per $\widehat{\mathrm{DP}}$ value in Fig. 5 (b) highlights how the output resolution shapes bit-wise decision errors. With a high output resolution, the MSE follows the INL shape in Fig. 4, such that quantization has negligible impact on the distribution of errors. However, decreasing the resolution discards increasingly more MSE contributions, eventually leading to decision errors only around $\widehat{\mathrm{DP}} = 0$ at a 1b output resolution. Hence, one should evaluate decision errors directly at the ADC output to account for the quantization shaping. Furthermore, the $\widehat{\mathrm{DP}}$ distribution will also affect these errors. Therefore, we introduce the following *distribution-aware* SNR metric between expected and actual digital outputs,

$$\mathrm{SNR} = \frac{\sum_{i=-\widehat{\mathrm{DP}}_{max}}^{\widehat{\mathrm{DP}}_{max}} p_i \sum_{j=0}^{N_{data,i}-1} \widehat{\mathrm{D}}_{out,i}^2}{\sum_{i=-\widehat{\mathrm{DP}}_{max}}^{\widehat{\mathrm{DP}}_{max}} p_i \sum_{j=0}^{N_{data,i}-1} \left(\widehat{\mathrm{D}}_{out,i} - \mathrm{D}_{out,(i,j)}\right)^2}. \quad (3)$$

$\widehat{\mathrm{DP}}_{max} = (2^{r_{in}} - 1)N_b$ is the maximum achievable $\widehat{\mathrm{DP}}$ and $N_{data,i}$ is the number of samples associated with the i-th

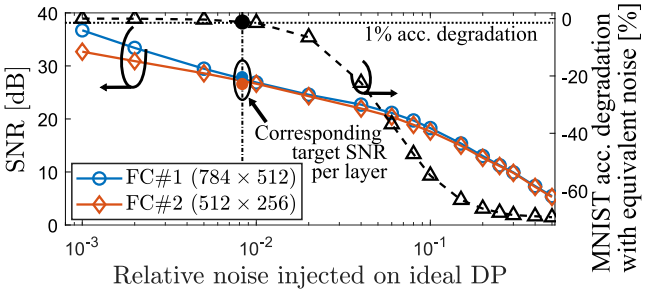

Fig. 6. Depicting the link between the accuracy degradation on MNIST and the SNR at the output of each layer of a $784 \times 512 \times 256 \times 10$ MLP, with binary inputs and weights. DP results are quantized to 4b.

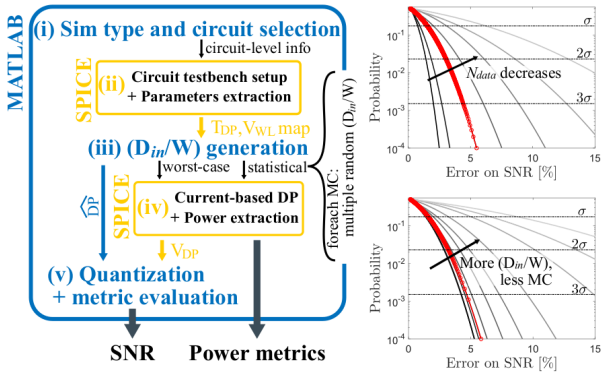

Fig. 7. MATLAB-SPICE simulation framework. For statistical simulations, we choose the sample size and repartition (red curves) to keep the 3σ simulation error on the SNR below 5% for the worst-case situation ($N_b = N_{rows}$).

DP value, with the total sample size summing up to N_{data}. We obtain each set of D_{out} samples by performing Monte-Carlo simulations on top of random (D_{in},W) combinations satisfying Eq. (1). Now, let us analyze the impact of the $\widehat{DP}$ distribution on the SNR. For a CNN topology with batch-normalizations [19] and (quantized) activations between successive CIM-SRAM layers, we assume a (discrete) gaussian distribution of the inputs. Furthermore, the weights distribution is highly dependent on the inference task, CNN architecture and layer position within the network [20]. Still, as we limit our scope to binary weights, we can assume they follow a binomial distribution $W \sim B(1, p_{bin})$. Fig. 5 (c) depicts the distribution for increasing p_{bin}, with uniformly-distributed weights corresponding to a value of 0.5. Then, we obtain the dependence of the SNR on the $\widehat{DP}$ distribution in Fig. 5 (d) when applying local mismatch. The asymmetry of the curves comes from the difference in signal power associated with negative and positive $\widehat{DP}$ values. At $p_{bin} = 0.5$, samples are mainly drawn around $\widehat{DP} = 0$, and the power increase with output resolution dominates. When p_{bin} approaches 1, the $\widehat{DP}$ distribution shifts towards higher values, requiring additional quantization thresholds to introduce new errors. Altogether, we conclude here that the SNR strongly relies on the $\widehat{DP}$ distribution. Nonetheless, we will fix p_{bin} to 0.5 throughout the rest of this work to simplify design space exploration.

In practice, *when working with a fixed network architecture and parameters*, one can link the accuracy degradation with the SNR at each layer output. We depict it in Fig. 6 when injecting a known level of equivalent noise on the ideal DP outputs. By extracting these noisy input/weights distribution to get actual D_{out} samples with our framework, one can check whether the acutal SNR resulting from Eq. (3) is larger than the target SNR in Fig. 6 for each layer. This guarantees to reach the targeted accuracy under the studied ANI constraints when mapping the network on the fixed CIM-SRAM macro.

E. Simulation Framework

Results presented in this work are based on SPICE circuit-level simulations embedded within a MATLAB framework, as depicted in Fig. 7. This framework operates in successive steps: (i) the simulation type and parameters, such as the supply voltage, are specified to SPICE, (ii) a first round of circuit-level simulations extracts critical parameters, in particular the maximum DP duration and the WL levels for multi-bit inputs. Stability simulations also occur at this stage, (iii) MATLAB generates the distribution of (D_{in},W) data, either for worst-case analyzes or statistical simulations, (iv) the corresponding WL inputs and bitcell data are fed to SPICE, which in turn extracts the analog DP result, (v) expected and actual results are quantized and, when desired, the SNR is finally derived. SNR evaluations always use the statistical distribution. Besides, note that power metrics are directly retrieved from SPICE.

For statistical simulations, we draw multiple (D_{in},W) combinations as described in the previous section. However, the choice of the total sampling size N_{data} sets a trade-off between computation speed and SNR accuracy. Moreover, for Monte-Carlo simulations, balancing this total sample size between Monte-Carlo runs and the number of random (D_{in},W) combinations per Monte-Carlo also affects accuracy. But, it allows to better distinguish mismatch *within* a single operator from mismatch *across* operators. We will take advantage of this distinction in Section VI. In order to keep the final SNR error below 5% for 3σ simulations, we take $N_{data} = 2 \times 10^4$, with 200 Monte-Carlo samples and 100 combinations per Monte-Carlo run when necessary. This sampling strategy corresponds to the highlighted red curves in Fig. 7 for the maximum operator size, which achieves the statistical worst-case simulation error.

III. Analysis and Mitigation of Systematic ANIs

The non-linearity of the access transistor current I_{on} is the backbone of systematic ANIs. Considering the α power law expression from [21], I_{on} evolves non-linearly with both the gate-to-source voltage $V_{gs} = V_{WL}(t) - V_{Q(B),0}$ and drain-to-source voltage $V_{ds} = V_{BL(B)} - V_{Q(B),0}$, with $V_{Q(B),0}$ the average bit-0 voltage level of accessed bitcells. Let us study how one can mitigate these non-linearities.

A. V_{ds} Non-Linearity Mitigation by T_{DP} Calibration

Let us first assume binary inputs. The drain-to-source modulation of I_{on} by the BL voltage integrates a non-linear error on the DP result over time. Hence, we should expect to reduce INL by decreasing the BL voltage swing, which is associated to the DP duration as seen in Section II.B. Besides, a shorter

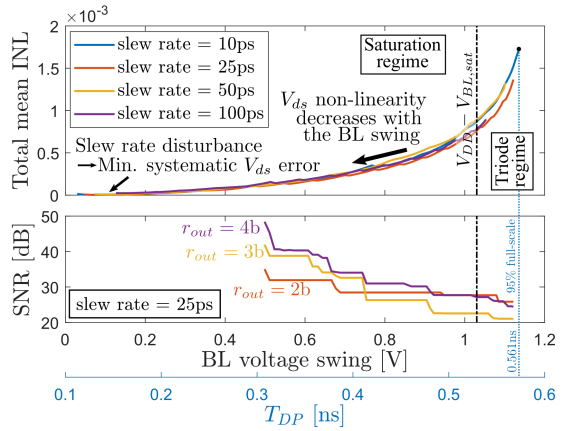

Fig. 8. Evolution of (top) the total mean INL of the DP result for different slew rates and (bottom) the SNR metric with the BL voltage swing and its corresponding DP duration, for $N_b = 16$ and increasing output resolution. Drain-to-source non-linearity decreases with a smaller BL swing.

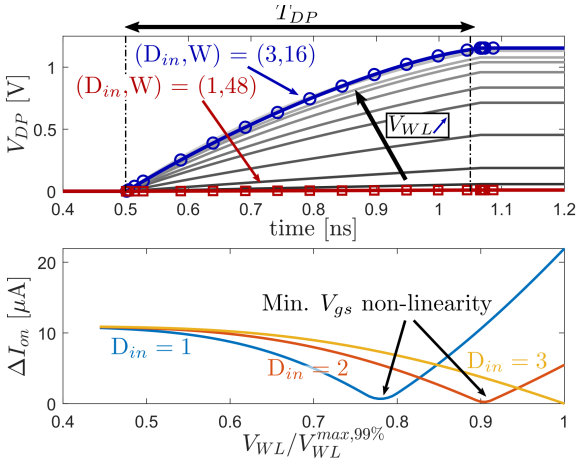

Fig. 9. (Top) Comparison of the actual DP result for (D$_{in}$,W) yielding an expected equivalent result. Tuning the WL voltage allows to fit the actual (D$_{in}$,W) = (1,48) result onto the (3,16) one. (Bottom) The non-linear WL DAC code is given by the abscissa of the minima of the ΔI_{on} curves.

DP duration would also decrease the integrated error from the total leakage current. We can evaluate this improvement by looking at the total mean INL across all $\widehat{\text{DP}}$s in Fig. 8, for different values of slew rate. The non-linearity reduction eventually stops when the slew rate becomes a large fraction of the DP duration. Observe that the INL decreases faster in the triode region because of exacerbated drain-to-source modulation associated with BL swing clipping. This suggests to keep the swing below the saturation limit $V_{DD} - V_{BL,sat}$, with $V_{BL,sat}$ the BL voltage at the edge of saturation. Minimum INL is obtained at very small BL swings. In practice, such swing would severely increase the sensitivity to noise and jitter while slowing ADC conversion. One would either need to waste ADC resolution bits or to adapt the ADC architecture for low-voltage [22], [23]. To prevent this, we fix the DP duration along the rest of this analysis so that the BL voltage swing reaches the saturation edge. Nonetheless, we shall review this trade-off in Section VI.

B. V_{gs} Non-Linearity Mitigation by WL Pre-Distortion

For multi-bit inputs, I_{on} is a non-linear function of the WL voltage levels which follow a linear, thermometer DAC code. Fig. 9 (top) illustrates this non-linearity for two (D$_{in}$,W) combinations, which give the same $\widehat{\text{DP}}$ result. While the actual DP results differ from the linear code, tuning the WL voltage level recovers the expected response. Hence, one could find the set of actual WL voltages V_{WL} that minimizes the integration error over the DP duration,

$$\Delta I_{on,i} = \int_{V_{BL,sat}}^{V_{DD}} \frac{|I_{on}(V_{WL}^{max,99\%}) - i\ I_{on}(V_{WL,i})|}{V_{DD} - V_{BL,sat}}\, \mathrm{d}V_{BL}.$$

Fig. 9 (bottom) shows the resulting error curves for a 2b input, with D$_{in}$ = 3 mapping to $V_{WL}^{max,99\%}$. Observe that the minimum error is non-zero because second-order effects exacerbated in short-channel devices correlate the dependencies on WL and BL voltages. Practically, achieving such programmable pre-distortion would require a dedicated DAC, similar to [24], but with much lower overheads by targeting an input resolution below 4b and by allowing some flexibility around the ΔI_{on} minima. As such, we shall consider the optimal mapping along the rest of this paper.

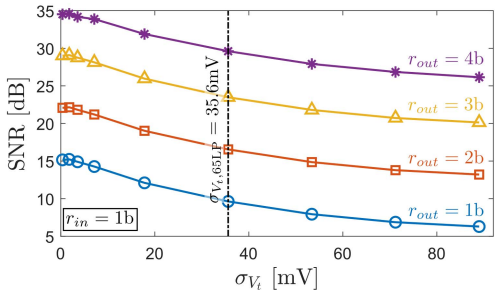

Fig. 10. Degradation of the SNR when increasing σ_{V_t} in 65nm LP CMOS (SVT) at 1.2V and $N_b = 256$, for different values of r_{out}. The actual V_t deviation for the technology reaches 35.6mV in these parametric conditions.

IV. Impact of Spatio-Temporal ANIs

With systematic ANIs addressed, let us compare the relative impact of each spatio-temporal ANI source highlighted in Fig. 2 (e) on the SNR. This comparison aims at providing a breakdown of the dominant ANI(s) in different corners of the design space.

A. Impact of Local Mismatch

Random dopant fluctuations and line-edge roughness distribute threshold voltages as $V_t \sim \mathcal{N}(\mu_{V_t}, \sigma^2_{V_t})$, which leads to strong local mismatch between each I_{on}. This high variability is usually pointed to as the main challenge in current-based CIM-SRAMs [17] and motivated the strive for other architectures [16]. Sweeping σ_{V_t} in baseline conditions and maximum operator size ($N_b = N_{rows}$) in Fig. 10 quickly reveals the SNR degradation. This degradation is quasi-linear (in dB scale) with σ_{V_t} once mismatch becomes substantial (here, $\sigma_{V_t} \gtrsim 10$mV) and shows a similar slope for all considered output resolutions. Indeed, most outputs are close to the 1b quantization threshold for the considered (D$_{in}$,W) distribution, yielding an identical increase in decision errors. In practice, σ_{V_t} is fixed by the technology and transistor type: we shall therefore consider

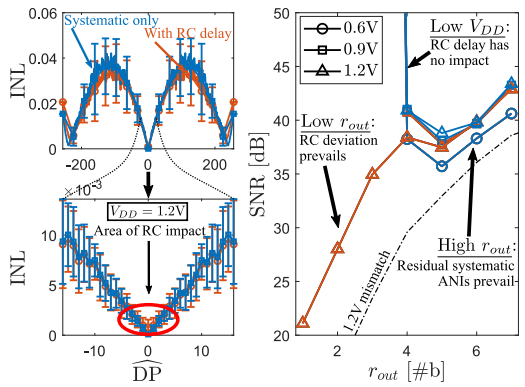

Fig. 11. (Left) Comparison of the INL on the DP result with and without R_{BL} parasitics for $N_b = 256$. The INL deviation is only altered at low $\widehat{DP}$ values. (Right) Impact of the $R_{BL}\,C_{BL}$ delay on the SNR, at different supply voltages. The SNR always remains above the reference mismatch level at $N_b = 256$.

the highlighted actual SNR for 65nm LP SVT transistors as an *upper bound* when comparing with other spatio-temporal ANIs in similar parametric conditions.

B. Impact of BL Parasitic Resistance

The parasitic resistance associated with the long and narrow BL metal wires create an $R_{BL}\,C_{BL}$ network, as illustrated in Fig. 2 (c). Because of the physical wiring distance between the ADC input and bitcells performing the DP operation, this network introduces *static position-dependent delays* leading to additional decision errors, assuming the DP duration is fixed. In the 65nm LP bulk CMOS technology, the estimation of the parasitic resistance for minimum-width 0.1μm metal-2 wires gives $R_{BL} = 0.91\Omega/\mu$m. We compare in Fig. 11 (left) the INL with the addition of these parasitics to the systematic-only case for the maximum operator size, which has the largest disparity in wiring distances. The INL deviation only rises around low $\widehat{DP}$ values, characterized by a large number of equivalent (D_{in},W) combinations, each with a different delay. This behavior explains that the additional SNR loss in Fig.11 (right) is only significant at a low output resolution, because one only accounts for decisions around $\widehat{DP} = 0$. Nonetheless, *the SNR remains high compared to the local mismatch reference in identical conditions* ($N_b = 256$ at 1.2V). Moreover, scaling the supply voltage down reduces the BL current density during the DP operation, which lowers the dynamic voltage drop across the parasitic R_{BL} and the resulting delay-based deviation of DP results. Eventually, we should underline that the small impact of these parasitics in 6T-based architectures is an advantage over existing single-ended designs that perform the DP as a voltage division on the bitline, such as [12]. While we only suffer here from a slight $R_{BL}\,C_{BL}$ delay, parasitic resistances in such designs introduce direct IR drops on the expected voltage division, which impacts the DC level of the DP result meaningfully.

C. Impact of IR Drops

Let us now assume N_{cells} pull-down paths act in parallel during a single DP operation, as represented in Fig. 12 (left). The peak current drawn from this large level of parallelism can induce severe IR drops across the parasitic wiring resistance to ground R_{gnd}. Assuming strong inversion, one finds the DC level of the virtual ground node V_{gnd}:

$$V_{gnd}^{DC} = R_{eq}\; I_{on} \text{ where } \begin{cases} R_{eq} & = R_{gnd}\; N_{cells}, \\ I_{on} & \propto (V_{WL} - V_{Q,0} - V_t)^2. \end{cases} \quad (4)$$

Changing the equivalent resistance in DC conditions with the BL voltage set to V_{DD} in Fig. 12 (right) gives the critical value for which IR drops start to significantly affect I_{on}, as the bit-0 node rises above its nominal read-access voltage level ($\simeq$ 70mV, from simulations). This critical value rises when one decreases the supply voltage, as the lower nominal read current induces smaller IR drops at fixed resistance value.

By sweeping V_{gnd}^{DC} in Fig. 13 (right), we find that IR drops quicky degrade the SNR, as they reduce both the actual BL voltage swing and I_{on}. Besides, the increased bit-0 voltage level also weakens bitcell stability and pushes the access transistors towards weaker inversion, making them more sensitive to mismatch. With the value of R_{gnd} obtained from layout parasitic extraction, one can deduce the maximum number of cells working in parallel from Fig. 12 (right) for a target SNR level in Fig. 13 by matching the values of V_{gnd}^{DC}. In that perspective, scaling the supply voltage down allows more parallelism, at the cost of a larger sensitivity to mismatch. Applying these considerations to the 65nm LP node, we obtain $R_{eq} = 780\Omega$ for our 256×64 array when assuming a meshed power grid (see Appendix B for the detailed computation). From Fig. 12, we conclude that such equivalent resistance does not impact V_{gnd}^{DC} meaningfully, leaving the SNR intact.

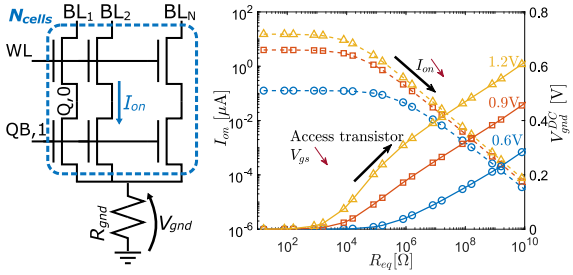

Fig. 12. (Left) Representation of IR drops affecting parallel DP pull-down paths. (Right) Concurrent evolution of I_{on} and V_{gnd}^{DC} with R_{eq} at different supply voltages. V_{gnd}^{DC} and thus $V_{Q,0}$ increase when R_{eq} increases, so that I_{on} starts decreasing when V_{gnd} overcomes the nominal value of $V_{Q,0}$.

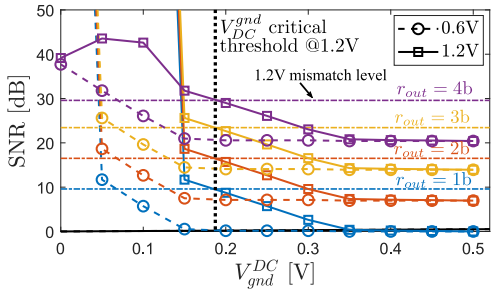

Fig. 13. Evolution of the SNR with V_{gnd}^{DC} at 1.2V and 0.6V. The SNR falls below the reference mismatch level above a critical value IR drops.

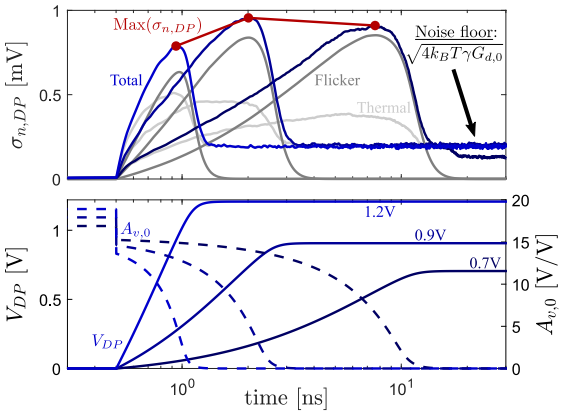

Fig. 14. Transient evolution of thermal, flicker and total RMS noise on V_{DP} during a DP operation, at different supply levels ($N_b = 16$, RMS averaged over 10^3 *noisetran* simulations). Noise power falls steeply off when the access transistors reach the edge of saturation, as $A_{v,0}$ collapses quickly.

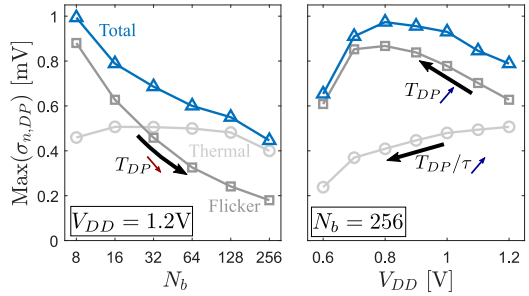

Fig. 15. Evolution of thermal, flicker and total maximum RMS voltage $\sigma_{n,DP}$ with (left) operator size and (right) supply voltage. The RMS noise stays below 1mV over both full parametric ranges.

D. Impact of Sample-and-Hold Noise

The accumulation of noise injected by bitcells on BL/BLB during DP operations behaves as a non-stationary stochastic process. Adapting the expressions of the thermal and flicker noise variances in [25] and [26] to our DP operator, we find

$$\begin{cases} \sigma^2_{n,th} = \dfrac{k_B T}{C_{BL}} \gamma \, \dfrac{G_m}{G_d} \left(1 - e^{-2t/\tau}\right) u(t) \text{ with } \tau = C_{BL}/G_d, \\ \sigma^2_{n,1/f} = \dfrac{G_m^2 k_F}{C_{ox} \text{WL}} \dfrac{t^2}{2\, C_{BL}^2} \Big(3/2 + \log(4T_s/t)\Big). \end{cases} \tag{5}$$

k_B is the Boltzmann constant, T is the temperature (in Kelvin), γ is the body-effect coefficient ($\simeq 2/3$, 1 for bulk technologies in saturation and triode, respectively [27]), G_m and G_d are respectively the total equivalent transconductance and output impedance seen by the bitline, k_F is a technology-specific factor, C_{ox} is the access transistor gate-oxide capacitance, and T_s is the characteristic time of the $1/f$ process. Given here that the DP duration is much smaller than T_s ($T_s \geq$ 1ms), flicker noise acts during several cycles as a static offset which depends quadratically on the DP duration. Furthermore, the total integrated thermal noise increases with the DP noise bandwidth $1/\tau$.

Knowing that each BL/BLB precharge resets the noise level, let us consider the transient response of this non-stationary process over an indefinitely long DP operation ($T_{DP} \in [0, +\infty[$). Fig. 14 gives the root mean square (RMS) value $\sigma_{n,DP}$ of the DP voltage due to these noise contributions

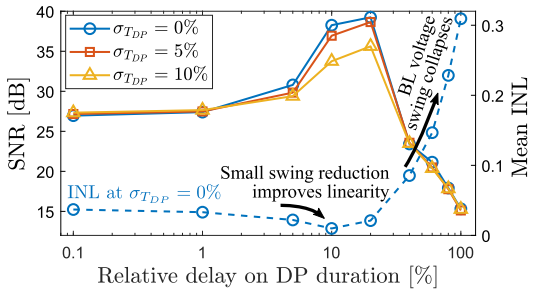

Fig. 16. Evolution of the SNR with WL delay at 1.2V and $N_b = 16$, with increasing relative jitter ($r_{out} =$ 4b). The INL improvement at moderate delay and without noise increases the SNR, before collapsing at larger ones.

over 10^3 noise realizations, in baseline conditions. The RMS voltage increases over time as predicted in Eq. (5), until the total gain $A_{v,0} (= G_m/G_d)$ collapses and transistors enter the triode region ($V_{BL} < V_{BL,sat}$). At that point, the flicker contribution disappears entirely and the RMS voltage settles around the thermal noise floor RMS at zero V_{ds}, equal to $\sqrt{4k_BT\gamma\, G_{d,0}}$ with $G_{d,0}$ the total off-mode impedance [27]. Hence, the integrated noise impact is maximum when V_{DP} is close to the saturation edge.

We further investigate in Fig. 15 how changing the operator size or supply voltage affects this maximum value. From Eq. (5), the shorter DP duration obtained when increasing the operator size reduces flicker noise, despite the linear increase of G_m with N_b. Moreover, the cancelling dependences on N_b in the opposing G_m/G_d and T_{DP}/τ ratios lead to bare fluctuations of the thermal noise. At first order, both ratios do however increase when scaling the supply voltage down, such that thermal noise decreases as in Fig. 15 (right). On the contrary, flicker noise increases with a longer DP duration, until it becomes close to T_s. Finally, advanced nodes will increasingly suffer from such noise because of BL capacitance reduction.

E. Impact of WL Timing Errors

Considering 1b inputs, timing errors on the WL signals also affect the DP result at fixed DP duration. Let us distinguish two kinds of timing errors.

1) Inter-Column Delay Error: The parasitic $R_{WL}\,C_{WL}$ network creates a horizontal WL propagation delay that shortens the actual pulse width seen at a given column, assuming the digitization of DP results by all column-ADCs occur simultaneously. Fig. 16 showcases the impact of the relative delay on the SNR at a 1.2V supply voltage and $N_b = 16$. The SNR begins to drop around a 20% relative delay and drops exponentially. Interestingly, we also observe an increase in SNR for moderate delays. In fact, such delay slightly improves the linearity of the transfer function, as verified with the INL curve. Still, we note that additional statistical variability shatter this improvement, such as the relative jitter $\sigma_{T_{DP}}$. Now, let us estimate the worst-case delay in the 65nm LP technology. With (i) 64 bitcells of width 1.05μm, (ii) the WL parasitic capacitance estimated to 25fF (half of C_{BL}), (iii) a parasitic resistance of 0.91Ω/μm for metal-3 wires, the delay of the 64-th column is only 4.28ps. Such value remains smaller than the 25ps slew rate, only affecting large operator sizes.

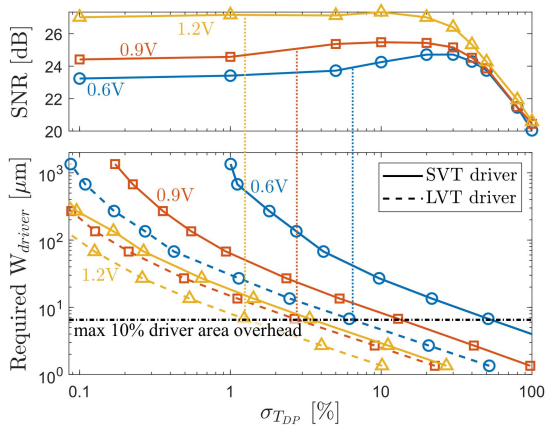

Fig. 17. (Top) Evolution of the SNR with normalized jitter $\sigma_{T_{DP}}$, for different supply voltages ($N_b = 16$, $r_{out} = 4$b). (Bottom) Correspondence between WL buffer sizing and relative WL jitter increase. Limiting the maximum buffer width sets the maximum jitter-induced SNR level.

2) Inter-Row Jitter: Noise and mismatch between WL drivers introduce an inter-row jitter with standard deviation $\sigma_{T_{DP}}$. Assuming a normal distribution of this jitter and neglecting noise, we evaluate its impact by applying random jitter to each WL on top of the (D$_{in}$,W) distribution. As such, we find in Fig. 17 (top) that the SNR starts to fall exponentially at a 20-30% relative jitter depending on the supply voltage, for 16 bitcells per operator. Yet, the actual jitter level depends on the WL driver strength. Let us consider the simple buffer topology supplied at $V_{WL}^{max,99\%}$ for 1b inputs. We can extract the actual jitter effectively introduced for any total width W$_{driver}$ of the WL driver and directly relate it to the SNR estimate, as done in Fig. 17. In that regard, low-V_t (LVT) devices enable to work with lower area to reach a same SNR. Assuming we allow up to 10% area overhead for the WL drivers compared to the total array, we find that LVT drivers ensure no SNR degradation with $N_b = 16$. Nonetheless, increasing the operator size would decrease the DP duration, requiring to upsize the WL driver to keep the SNR constant.

For multi-bit inputs, mismatch would affect both the pulse width and the generated WL voltage level. In order to avoid a statistical V_{gs} non-linearity, one should size the DAC to keep the WL voltage levels close to the minima in Fig. 9. As for the 1b WL driver, this requirement introduces a trade-off between area and accuracy, but also depends upon the resilience of the selected DAC architecture to noise and mismatch.

F. SNR Breakdown

Finally, we can dress a breakdown of the SNR under the typical hardware conditions described for each ANI. Table I showcases that *local mismatch prevails in most corners of the design space*. Observe that WL timing errors become critical as well at large operator size due to the 10% area overhead constraint on the WL driver. However, relaxing this constraint could directly improve the SNR back.

V. Analysis of Design Space Trade-Offs

Considering that local mismatch is the only substantial source of non-systematic SNR degradation, let us consider how the trade-off between SNR, power metrics and throughput evolve across the design space of the 6T current-DP operator. Moreover, area efficiency is obtained as throughput/area, with the 65nm LP 6T bitcell area of 0.5μm^2 with dense rules and about 1μm^2 with logic rules, here considered [16]. Note that power and throughput *are only extracted for the bitcell array, yielding upper bounds results unconstrained by the WL/BL drivers and periphery*. We speak hereafter of *unbounded* power and throughput. For power metrics, we evaluate both the total leakage power normalized per bitcell (P_{leak}) and the power efficiency (PE). Power efficiency is estimated in the worst-case scenario where all WLs and BLs/BLBs are charged to full-swing. For *interpreting* the SPICE-extracted power figures, let us use the following analytical expressions,

TABLE I
SNR Breakdown in Typical Conditions

	0.6V, $N_b = 16$		0.6V, $N_b = 256$		1.2V, $N_b = 256$	
r_{out}	1b	4b	1b	4b	1b	4b
Systematic only	$+\infty$	24.5dB	$+\infty$	39.7dB	$+\infty$	40.9dB
Mismatch	4.7dB	13.6dB	3.9dB	22.5dB	9.6dB	29.3dB
$R_{BL}\,C_{BL}$ delay	$+\infty$	24.3dB	$+\infty$	38.2dB	21.1dB	38.5dB
IR drops	$+\infty$	21.6dB	$+\infty$	39.2dB	$+\infty$	40.6dB
Intrinsic BL noise	$+\infty$	18.3dB	39.6dB	33.4dB	$+\infty$	39.8dB
WL timing error	13.5dB	23.3dB	5.8dB	27.2dB	9.5dB	33.3dB

$$\begin{cases} P_{leak} = V_{DD}\, I_{cell}(V_{DD}) \text{ and } \mathrm{PE} = N_{op}/E_{DP}, \text{ where} \\ E_{DP} \simeq (2\, N_{cols} C_{BL} + N_b C_{WL})\, V_{DD}^2 \\ \qquad + N_{rows} N_{cols} \int_0^{T_{DP}} P_{leak}\, \mathrm{d}t. \end{cases} \tag{6}$$

I_{cell} is the current drawn from supply by a single bitcell (for static leakage, we assume $V_{WL} = V_{BL(B)} = 0$V), $N_{op} = 2 \times N_b \times N_{cols}$ is the number of addition *and* multiplication computed by all of the bitcells involved in parallel DP operations. Moreover, the throughput is equal to N_{op}/T_{DP}. Let us underline that leakage power is often overlooked in the CIM-SRAM literature compared to power efficiency, although this metric is key when targeting edge applications with sparse activity, especially with volatile memories like SRAMs [28].

A. Transistor Type

Memory designers can select amongst several process flavors and transistor types to trade-off read/write latency, bitcell leakage and area [29]. For example, in 65nm CMOS, one usually finds a Low-Power (LP) and a General-Purpose (GP) process flavor, with different V_t's, printed gate length and effective gate-oxide thickness [30]. Note that these flavors use here different nominal supply voltages: 1V for GP and 1.2V for LP. The differences in sensitivity to V_t variations amongst these flavors is critical for mismatch resilience. In light of this, Fig. 18 (top) showcases a general increase in V_t deviation when moving from GP to LP flavors, or from low-V_t (LVT) to high-V_t (HVT) devices, with the exception of the GP LVT one. The relative I_{on} variability $(\sigma/\mu)_{I_{on}}$ follows the V_t deviation trend, with higher sensitivity at large V_t as transistors are biased closer to weak inversion, at constant V_{gs}.

We derive the trade-off between SNR and power metrics in Fig. 18 (bottom). It clearly highlights the trade-off between low leakage power at high V_t and high SNR at low V_t

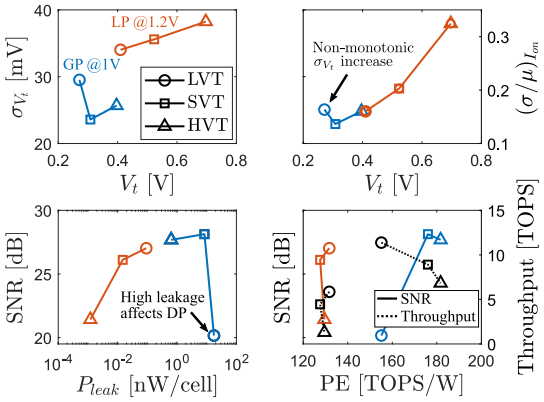

Fig. 18. (Top) Comparison of sensitivities to V_t (left) and I_{on} (right) changes across all transistor types in the 65nm bulk CMOS technology. (Bottom) Trade-off between SNR ($r_{out} = 4$b) and (left) bitcell leakage power, (right) power efficiency over 64 columns.

thanks to the lower I_{on} variability. GP LVT devices are an exception because their high leakage cannot be neglected compared to I_{on}, hence noticeably distorting the DP result. Regarding power efficiency, LP devices have a negligible leakage contribution, such that the BL/BLB precharge cost dominates. We note slight variations with their V_t type, related to changes in the total intrinsic drain capacitance connected to BL/BLB. Still, GP devices showcase a better power efficiency thanks to their reduced nominal supply voltage. However, their high total leakage quickly degrades this efficiency when moving from HVT to LVT. These devices also have a better throughput because of their lower V_t, in spite of the supply voltage difference. Altogether, we conclude here to select GP devices to reach high power efficiency *and high SNR*, and LP devices for applications requiring low standby power.

B. Technology and Supply Voltage Scaling

While technology scaling increases the CIM-SRAM density, it also worsens V_t variability following Pelgrom's law [31],

$$\sigma_{V_t} = A_{V_t}/\sqrt{\mathrm{W\,L}}. \tag{7}$$

In Eq. (7), W and L are the transistor width and length, and A_{V_t} is a technology-dependent parameter. Table II compares the technology data of increasingly scaled nodes. For bulk technologies, we observe that A_{V_t} improvement with scaling can generally not follow the reduction of the $\sqrt{\mathrm{W\,L}}$ factor, leading to an increased V_t variability. We should also note that FD-SOI technologies have much lower A_{V_t} than bulk CMOS processes thanks to inherently undoped channels [32], explaining the lower V_t deviation of the 28nm FD-SOI node. Let us now investigate how the SNR associated with these technologies evolves with supply voltage scaling. Fig. 19 (top) presents the evolution of the relative I_{on} variability when scaling it down to half its nominal value for each technology. Each value is extracted at the corresponding $V_{WL}^{max,99\%}$ voltage, which decreases quasi-linearly with the supply voltage to keep the driving strength ratio between pull-up PMOS and access transistor constant. Hence, the relative I_{on} variability rises due to the lower WL voltage, quickly pushing the access transistors

TABLE II
LIST OF TECHNOLOGY-SPECIFIC PARAMETERS

	V_{DD}^{nom}* [V]	$(\mathrm{W/L})_{min}$* [nm/nm]	V_t† [V]	A_{V_t}† [nm·V]	σ_{V_t}† [mV]	C_{BL}‡ [fF]
0.18μm bulk CMOS	1.8	240/180	0.39	3.45	18.1	120
0.13μm GP SVT bulk CMOS	1.2	150/130	0.29	3.57	25.2	90
65nm GP SVT bulk CMOS	1	135/60	0.31	2.14	23.6	50
65nm LP SVT bulk CMOS	1.2	135/60	0.52	3.19	35.6	50
28nm SVT FD-SOI	0.9	80/30	0.33	0.95	17.3	25

* From technology datasheets.
† From SPICE simulations.
‡ From post-layout extraction (M2).

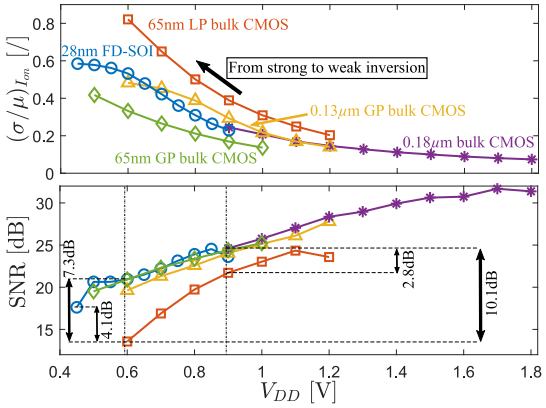

Fig. 19. Evolution of (top) the normalized I_{on} deviation and (b) the SNR ($r_{out} = 4$b, $N_b = 16$) with supply voltage scaling for different technologies. At constant supply, the lower normalized current deviation achieves the best SNR. The 65nm LP technology has the worst SNR because of its higher V_t.

towards weak inversion. The steep slope in Fig. 19 (top) characterizes this transition from strong to weak inversion. Such behavior strongly degrades the SNR, as seen in Fig. 19 (bottom) for a 4b output resolution. Comparing technologies at fixed supply voltage, a lower relative I_{on} variability means a better SNR because the mismatch contribution dominates that of systematic ANIs, except for very close values of $(\sigma/\mu)_{I_{on}}$.

Let us now discuss the SNR, power and throughput trade-offs facing these scaling aspects in Fig. 20. On the one hand, leakage power increases exponentially with the supply voltage as the bitcell inverters are biased in weak inversion. On the other hand, there is a clear trade-off between power efficiency and SNR/maximum throughput as we scale the supply voltage down. Observe that the power efficiency begins to saturate for the 28nm FD-SOI and 65nm GP technology at low supply voltage: DP duration increases, so that the total integrated leakage eventually overcomes the diminishing WL/BL precharge costs. Furthermore, technology scaling reduces the total WL and BL capacitances, which improves power efficiency at a fixed supply voltage. Overall, the FD-SOI technology shows excellent trade-offs between SNR, power metrics and throughput at fixed supply voltage, leading us to the conclusion that FD-SOI designs offer promising perspectives for the further scaling of current-based CIM-SRAMs.

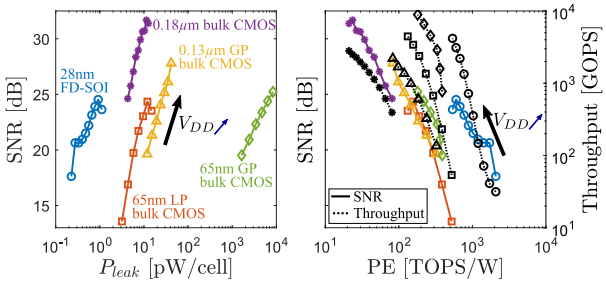

Fig. 20. Trade-off between SNR (r_{out} = 4b, N_b = 16) and (left) bitcell leakage power and (b) power efficiency/throughput with supply voltage scaling, for different technologies. FD-SOI technology provides the lowest leakage thanks to substrate isolation, and the highest power efficiency because of reduced supply and WL/BL capacitance.

C. Operator Size

The ability to scale the operator size and/or the input resolution in CIM-SRAM designs is key to enable the mapping of different CNN topologies. Yet, tuning these parameters affects the SNR as they change the range of $\widehat{\text{DP}}$ results and the number of equivalent (D_{in},W) combinations associated with each result. Moreover, while changing the operator size does not affect leakage power as the number of rows is fixed, Eq. (6) predicts a linear increase in power efficiency with N_b if

$$2\,C_{BL} \gg (N_b/N_{cols})\,C_{WL} \tag{8}$$

when neglecting the impact of the integrated leakage power. As a corollary, CIM-SRAMs with low aspect ratio N_{rows}/N_{cols} do not improve in power efficiency when expanding the operator size. Given that Eq. (8) yields $C_{WL} \leq 1/9\,C_{BL}$ in our baseline framework, this assumption is partially correct.

Fig. 21 quantifies the dependences on N_b (exponentially spanned from 8 to 256 bitcells) and r_{in} for different output resolutions, assuming other baseline conditions (1.2V, 65nm LP SVT). Fig. 21 (right) validates the linear increase in power efficiency with the operator size, and shows a supra-linear increase in the unbounded throughput. However, remember from Section IV.E that the tiny DP duration at large operator size makes the DP very sensitive to WL timing errors at fixed WL driver width. More interestingly, this figure showcases that the operator size that maximizes the SNR increases with the output resolution. To understand this and ease the simultaneous vizualization of different N_b curves, let us define $\widehat{\text{DP}}_{eq} = (N_{rows}/N_b) \times \widehat{\text{DP}}$. Fig. 22 demonstrates that two opposing effects lead to such behavior. On the one hand, Fig. 22 (top) shows an increase of $\text{INL}_{eq} = (N_b/N_{rows}) \times \text{INL}$, which corresponds to more quantization errors. On the other hand, the $\widehat{\text{DP}}_{eq}$ distribution concentrates around $\widehat{\text{DP}}_{eq} = 0$ when the operator size increases. For a 1b output resolution, a large number of data samples are located around the sole 1b quantization level, whatever the operator size. Therefore, the increasing INL_{eq} results in a larger quantization error, decreasing the SNR accordingly. However, for multi-bit outputs, we see that the number of data samples around additional quantization levels shrinks with N_b. Consequently, these samples do not provide additional quantization errors on the output. As a result, the SNR improves despite the larger

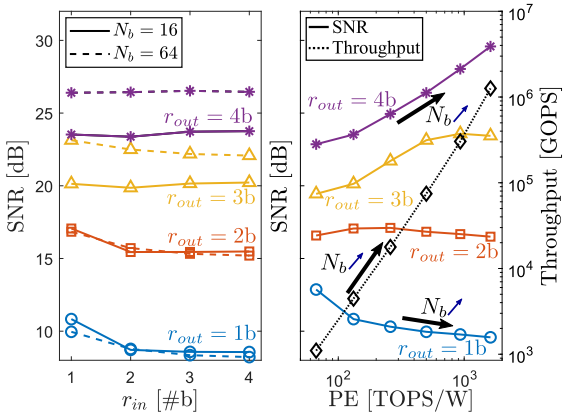

Fig. 21. Evolution of the SNR (65nm LP at 1.2V) with (left) r_{in} for different output resolutions, and (right) N_b (from 8 to 256) with the corresponding throughput and power efficiency. The operator size which maximizes the SNR-PE trade-off increases with the output resolution.

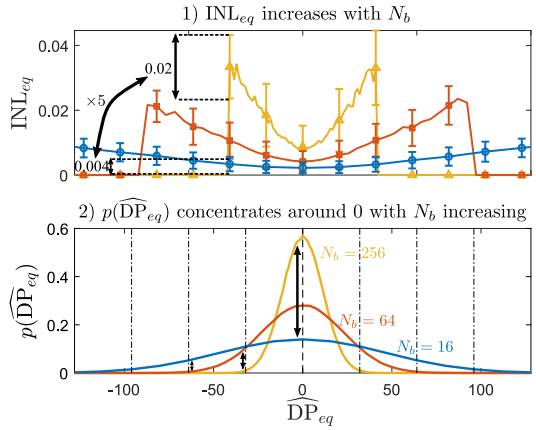

Fig. 22. Evolution of (top) INL_{eq}, (bottom) the $\widehat{\text{DP}}_{eq}$ distribution with N_b at 1.2V. These phenomena oppose one-another and result in an increase of the operator size maximizing the SNR with r_{out}.

INL_{eq} when increasing the output resolution. The maximum SNR is reached when the sample size around all non-binary quantization levels becomes close to zero, such that the INL_{eq} increase prevails again.

Eventually, we find in Fig. 21 (left) that the SNR degrades with the input resolution at low output resolution and fixed operator size. The reason is twofold: first, the effective number of steps increases with r_{in} as $\widehat{\text{DP}}_{max} = (2^{r_{in}} - 1)\,N_b$. Given that the DP duration remains unchanged, fitting more steps within the same voltage range increases INL_{eq}. Second, using a DAC makes inputs with reduced WL voltage more sensitive to mismatch. Nevertheless, we only notice the SNR decrease when moving from 1b to 2b input resolution. Besides, statistical averaging amortizes the drop at larger operator size and output resolution.

In the end, we conclude that striving concurrently for a high operator size and output resolution gives the best SNR and power efficiency results. Besides, it minimizes the SNR sensitivity to changes in input resolution. However, we should underline that this conclusion remains highly dependent on the $\widehat{\text{DP}}$ distribution.

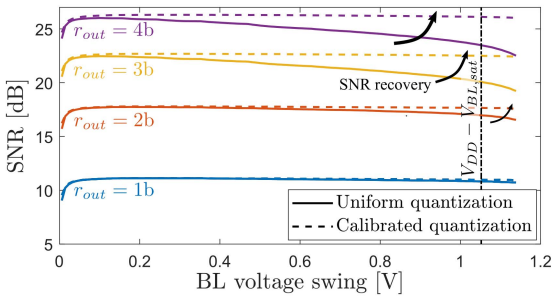

Fig. 23. Evolution of the SNR with the maximum BL voltage swing, for uniform and calibrated ADC quantizations (baseline conditions). The SNR loss at high output resolution driven by non-linear effects is mostly recovered through calibration.

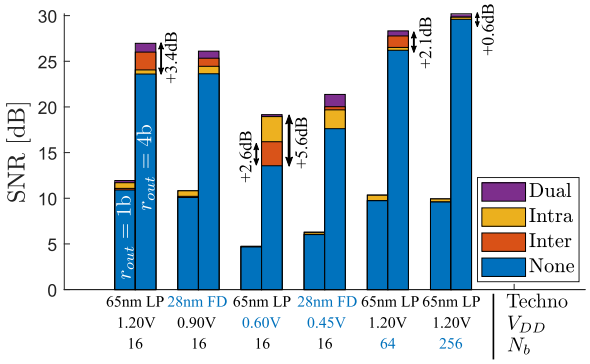

Fig. 24. Comparison of SNR improvements using calibration techniques, for different corners of the design space (SVT transistors). These techniques showcase a better effectiveness at high r_{out}, low V_{DD} and small N_b.

VI. Pushing the Design Space Limits

With the design trade-offs of 6T-based current-DP operators exposed, let us finally consider how hardware calibration techniques could alleviate the design constraints by improving resilience to mismatch and non-linearity, in particular.

A. Strategies of Statistical Calibration

Taking a step back, the overall *inter-operator* distribution of DP results comes from the accumulation of individual *intra-operator* responses. These intra-operator transfer functions suffer from identical systematic errors but different local mismatches. As calibrating each bitcell is not possible in practice, let us rather consider distribution-aware calibration techniques respectively acting at the inter- and intra-operator levels.

1) Inter-Operator Distribution-Aware ADC Quantization: Yin *et al.* proposed to manually tune the quantization levels of custom flash ADCs by using external supplies, so as to account for the actual distribution of outputs extracted from chip measurements [12]. However, we could instead use the knowledge of the statistical inter-operator distribution in order to infer, at design time, the calibrated ADC quantization levels which maximize the SNR. This solution bypasses the need for external supplies, as we could directly specify the quantization levels by software-configured registers of the CIM-SRAM. These registers could control a dedicated DAC that periodically refreshes the $2^{r_{out}}$ non-linear decision thresholds of r_{out}-bit flash ADCs.

2) Intra-Operator Gain and Offset Compensation: Assuming the choice of BL voltage swing makes the nominal transfer function almost linear, we could reduce the actual deviation of the inter-operator distribution by aligning the transfer functions of each operator on average by applying gain and offset compensations γ_n and β_n specific to each column, such that

$$V_{cal,n} = \gamma_n V_{DP,n} + \beta_n, \text{ with } n \in [0, Ncols - 1]. \quad (9)$$

one can find each γ and β by simply estimating the DP voltage obtained for the maximum and minimum DP result in each operator, as these results are inherent averages of the N_b currents respectively on the BL and BLB side. With regards to practical implementation, tuning of the BL capacitance (e.g. with a digitally-controlled capacitive bank) or column-wise back-biasing (for FD-SOI technology [33]) could apply the gain corrections γ_n. Additional calibration bitcells could generate the β_n factors by applying an offset on each $V_{DP,n}$. However, technology scaling degrades the efficiency of such a calibration as the V_{ds} non-linearity in saturation increases.

B. SNR Improvement Results

With these calibration techniques defined, let us first take a look back at the BL voltage swing selection with the addition of local mismatch, as announced in Section III. With uniform quantization, the SNR level in Fig. 23 (including now local mismatch) degrades more steeply with a large output resolution, as the sensitivity to non-linearity rises. However, we observe that this dwindling fades away with the proposed inter-operator calibration, which takes the stochastic non-linearity of the outputs into consideration. Therefrom, we recover one degree of freedom on the BL swing selection at design stage, reducing it to a trade-off between noise, speed and power.

Still assuming the initial BL swing, Fig. 24 quantifies SNR improvements using the proposed calibration for different sets of design space parameters. The SNR recovery is more effective at high output resolution as inter-operator calibration also tackles the residual non-linearity. Besides, we find that intra-operator calibration induces a better recovery of the SNR at low supply voltage, related to a better linearity of the transfer function in weak inversion. On the contrary, increasing the operator size limits the effectiveness of the iterative inter-operator calibration because the importance of non-binary quantization levels diminishes. Also, we limited the voltage resolution when calibrating the quantization levels to 2% of the supply voltage. Finally, applying dual calibration (i.e. ADC calibration applied on the aligned intra-operator distributions) always yields the best SNR improvement. Altogether, these techniques alleviate some design space trade-offs: they especially enable to work at lower supply voltage for a fixed target SNR, improving concurrently leakage power and efficiency.

Nevertheless, the overhead and limitations associated with the actual implementation of these techniques need a careful study. Let us give here a coarse overview of the power/area overheads. In the 65nm LP technology, eight 8-bit HVT registers based on standard-cell flip-flops would generate 1.5%

additional leakage power and 9.7% area overhead (compared to the bitcell array only). These look acceptable if the 8-bit DAC generating the quantization levels fits within the WL drivers pitch and consumes a low average power. For intra-operator calibration with 4b resolution on γ and β, one could use 16 bitcells and a 4b gain correction circuit. While the bitcells bring 6.25% static power and area overheads, the gain correction circuit would need a finer design attention.

VII. Conclusion

In this paper, we addressed the impact of analog non-idealities (ANIs) on the design space of analog dot-product (DP) operators based on 6T-SRAM bitcells for in-memory computing. We introduced the distribution-aware SNR metric to objectively quantify bit-errors on the DP results and showed that assumptions taken on the distribution of the DP operands critically affect the obtained error level. We proposed solutions to systematic ANIs by carefully analyzing their sources. For the 65nm LP node, we found out that parasitic RC delays and intrinsic noise have negligible impact on the SNR compared to local mismatch. Furthermore, we detailed how one can prevent IR drops and WL jitter from impacting the SNR by respectively finding the maximum number of parallel DPs and appropriately sizing the WL drivers. Assuming these conditions, we quantified the design trade-offs between the mismatch-induced SNR degradation and power consumption metrics. The scaling of supply voltage and V_t with CMOS technology usually improves power efficiency at the cost of more leakage and current variability, which degrades the SNR. Instead, we showed that FD-SOI technologies provide a better scaling solution thanks to inherently lower leakage and V_t mismatch. Besides, maximizing concurrently the operator size and output resolution improves at the same time the power efficiency, throughput and SNR. Finally, we discussed inter- and intra-operator calibration techniques enabling a partial recovery of the lost SNR. They especially allow to scale supply voltage down while preserving the same SNR level, hence improving power metrics. Altogether, this work provides a physical and quantitative roadmap of the main design-stage trade-offs faced by current-domain DP operators based on 6T-SRAM bitcells. The proposed framework with its distribution-aware metric also paves the way for applying such analysis to other in-memory DP operators.

Appendix A
Summary of Symbol Definitions

Table III regroups the definitions of recurrent symbols used in this work. Symbols defined and used within a single subsection are not displayed here, as well as standard symbols and abreviations in electronics (V_{DD}, INL, ...).

Appendix B
Estimation of R_{eq}

When estimating R_{eq} for the proposed CIM-SRAM framework, we assume a meshed power grid layouted in M4/M5 metal layers and surrounded by an ideal power ring, as shown in Fig. 25. M4 to M5 vias ensure the connection between horizontal and vertical power lines, which are 0.3μm wide (W_m) in the considered 65nm node. In this technology, the sheet resistance R_{sh} for Mx (x $\geq$ 2) layers is equal to $0.12\Omega/\square$ at this width. Besides, we consider a $0.5\mu\text{m}^2$ area (A) for the foundry 6T bitcell, with a 2.39 width/height aspect ratio (AR).

TABLE III
Definition of Recurrent Symbols

Category	Symbol	Definition
General	ANI(s)	Analog non-ideality(-ies)
	DP	Dot-product
Architecture-related	BL, V_{BL}, C_{BL}	Bitline, bitline voltage and total capacitance
	WL, V_{WL}, C_{WL}	Wordline, wordline voltage and total capacitance
	N_{rows}, N_{cols}	Number of rows and columns in the CIM-SRAM framework
	$V_{Q,0/1}$	Analog voltage level of the stored bit-0/1 data
	V_{dsat}, $V_{BL,sat}$	Drain-to-source saturation voltage and the corresponding BL voltage
Dot-product symbols	D_{in}, D_{out}	Fixed-point input and output data to the CIM-SRAM
	$\widehat{\text{D}}_{out}$	Expected fixed-point output
	V_{DP}	Differential voltage representing the analog DP result
	$\widehat{\text{DP}}$	Expected floating-point dot-product result
	W	+1/-1 weights, mapped to 1/0 and 0/1 bit values and stored in bitcells (Note: sometimes used for the transistor width, see context)
Dot-product operation	T_{DP}	Duration of the analog DP operation
	I_{on}	Current drawn by the bit-0 side of one accessed bitcell
	$V_{WL}^{max,99\%}$	Maximum WL voltage for 99% CIM-SRAMs without any DP bit-flip
Analysis parameters	r_{in}, r_{out}	Input and output resolution
	N_b	Size (i.e. number of bitcells) of the analog DP operator ($\leq N_{rows}$)

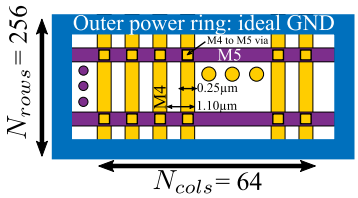

Fig. 25. Meshed power grid considered when estimating R_{eq}.

Because of the rectangular power ring, bitcells at the center of the array see the largest R_{gnd}. Note that we take this worst-case value as a constant in Eq. 4, therefore overestimating the actual level of IR drops. Given the previous set of hypotheses, we find the equivalent resistance R_{eq} as

$$R_{eq} \simeq N_{cells}\left(\frac{1}{R_v}+\frac{1}{R_h}\right)^{-1}$$
$$\text{with}\begin{cases} R_v = \dfrac{R_{sh}}{2}\dfrac{N_{rows}\sqrt{\text{A}/\text{AR}}}{N_{col}\,W_m}, \\ R_h = \dfrac{R_{sh}}{2}\dfrac{N_{col}\sqrt{\text{A}\times\text{AR}}}{N_{rows}\,W_m}. \end{cases}$$

The factor 1/2 comes from the two-sided connection to the ideal power ring. Using the previous data and $N_{cells} = 256 \times 64$, we eventually obtain $R_{eq} = 780\Omega$ in 65nm CMOS.

Acknowledgment

The authors would like to thank S. Cosemans for sharing his experience in CIM-SRAM design and for his precious advices toward the elaboration of this work.

References

[1] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.
[2] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.
[3] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems," *Proc. IEEE*, vol. 103, no. 8, pp. 1379–1397, Aug. 2015.

[4] Q. Dong *et al.*, "A 351 TOPS/W and 372.4 GOPS compute-in-memory SRAM macro in 7 nm FinFET CMOS for machine-learning applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 242–244.

[5] J.-W. Su *et al.*, "A 28 nm 64 Kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 240–242.

[6] V. Joshi *et al.*, "Accurate deep neural network inference using computational phase-change memory," *Nature Commun.*, vol. 11, no. 1, p. 2473, Dec. 2020.

[7] H. Jia, H. Valavi, Y. Tang, J. Zhang, and N. Verma, "A programmable heterogeneous microprocessor based on bit-scalable in-memory computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 9, pp. 2609–2621, Sep. 2020.

[8] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to $+1$ or -1," 2016, *arXiv:1602.02830*. [Online]. Available: http://arxiv.org/abs/1602.02830

[9] M. Kang *et al.*, *Deep In-Memory Architectures for Machine Learning*. Cham, Switzerland: Springer, 2020.

[10] A. Jaiswal, I. Chakraborty, A. Agrawal, and K. Roy, "8T SRAM cell as a multibit dot-product engine for beyond von Neumann computing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 11, pp. 2556–2567, Nov. 2019.

[11] M. Ali, A. Jaiswal, S. Kodge, A. Agrawal, I. Chakraborty, and K. Roy, "IMAC: In-memory multi-bit multiplication and accumulation in 6T SRAM array," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 8, pp. 2521–2531, Aug. 2020.

[12] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," *IEEE J. Solid-State Circuits*, vol. 55, no. 6, pp. 1733–1743, Jun. 2020.

[13] W.-H. Chen *et al.*, "A 65 nm 1 Mb nonvolatile computing-in-memory ReRAM macro with sub-16 ns multiply-and-accumulate for binary DNN AI edge processors," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 494–496.

[14] X. Peng *et al.*, "DNN+NeuroSim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies," in *IEDM Tech. Dig.*, Dec. 2019, pp. 32.5.1–32.5.4.

[15] X. Si *et al.*, "A 28 nm 64 Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 246–248.

[16] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A 64-tile 2.4-mb in-memory-computing CNN accelerator employing charge-domain compute," *IEEE J. Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, Jun. 2019.

[17] J. Zhang, Z. Wang, and N. Verma, "A machine-learning classifier implemented in a standard 6T SRAM array," in *Proc. IEEE Symp. VLSI Circuits (VLSI-Circuits)*, Jun. 2016, pp. 1–2.

[18] T. Haine, J. Segers, D. Flandre, and D. Bol, "Gradient importance sampling: An efficient statistical extraction methodology of high-sigma SRAM dynamic characteristics," in *Proc. Design, Automat. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 195–200.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[20] I. Bellido and E. Fiesler, "Do backpropagation trained neural networks have normal weight distributions?" in *Proc. ICANN*, S. Gielen and B. Kappen, Eds. London, U.K.: Springer, 1993, pp. 772–775.

[21] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr. 1990.

[22] J. E. Proesel and L. T. Pileggi, "A 0.6-to-1 V inverter-based 5-bit flash ADC in 90 nm digital CMOS," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2008, pp. 153–156.

[23] D. Daly and A. Chandrakasan, "A 6-bit, 0.2 V to 0.9 V highly digital flash ADC with comparator redundancy," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 3030–3038, Nov. 2009.

[24] J.-C. Pena-Ramos, K. Badami, S. Lauwereins, and M. Verhelst, "A fully configurable non-linear mixed-signal interface for multi-sensor analytics," *IEEE J. Solid-State Circuits*, vol. 53, no. 11, pp. 3140–3149, Nov. 2018.

[25] T. Sepke, P. Holloway, C. G. Sodini, and H.-S. Lee, "Noise analysis for comparator-based circuits," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 3, pp. 541–553, Mar. 2009.

[26] M. A. Ghanad, C. Dehollain, and M. M. Green, "Noise analysis for time-domain circuits," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 149–152.

[27] B. Wang, J. R. Hellums, and C. G. Sodini, "MOSFET thermal noise modeling for analog integrated circuits," *IEEE J. Solid-State Circuits*, vol. 29, no. 7, pp. 833–835, Jul. 1994.

[28] D. Bol *et al.*, "A 40-to-80 MHz sub-4 μW/MHz ULV cortex-M0 MCU SoC in 28 nm FDSOI with dual-loop adaptive back-bias generator for 20 μs wake-up from deep fully retentive sleep mode," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 322–324.

[29] J. Lee and A. Davoodi, "Comparison of dual-Vt configurations of SRAM cell process-induced Vt," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2007, pp. 1–4.

[30] D. Bol, R. Ambroise, D. Flandre, and J.-D. Legat, "Interests and limitations of technology scaling for subthreshold logic," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 10, pp. 1508–1519, Oct. 2009.

[31] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[32] O. Weber *et al.*, "High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding," in *IEDM Tech. Dig.*, Dec. 2008, pp. 1–4.

[33] G. de Streel and D. Bol, "Study of back biasing schemes for ULV logic from the gate level to the IP level," *J. Low Power Electron. Appl.*, vol. 4, no. 3, pp. 168–187, Jul. 2014.

Adrian Kneip (Student Member, IEEE) received the M.Sc. degree in electrical engineering from the Université Catholique de Louvain (UCLouvain), Louvain-la-Neuve, Belgium, in 2019, where he is currently pursuing the Ph.D. degree with the Electronics Circuits and Systems Group. His research interests include the modeling and design of mixed-signal ultra-low power systems and machine-learning hardware, with special dedication for SRAM memories, and in-memory computing.

David Bol (Senior Member, IEEE) received the M.Sc. degree in electromechanical engineering and the Ph.D. degree in engineering science from Université Catholique de Louvain (UCLouvain), Louvain-la-Neuve, Belgium, in 2004 and 2008, respectively. In 2005, he was a visiting Ph.D. student with the CNM National Centre for Microelectronics, Sevilla, Spain, in advanced logic design. In 2009, he was a Postdoctoral Researcher with intoPIX, Louvain-la-Neuve, Belgium, in low-power design for JPEG2000 image processing. In 2010, he was a Visiting Postdoctoral Researcher with the UC Berkeley Laboratory for Manufacturing and Sustainability, Berkeley, CA, in life-cycle assessment of the semiconductor environmental impact. He is currently an Associate Professor with UCLouvain. In 2015, he participated to the creation of e-peas semiconductors, Louvain-la-Neuve, Belgium. He leads the Electronic Circuits and Systems (ECS) group focused on ultra-low-power design of integrated circuits for the IoT and biomedical applications including computing, power management, sensing and wireless communications. He is engaged in a social–ecological transition in the field of ICT research. He co-teaches four M.S. courses in electrical engineering at UCLouvain on digital, analog and mixed-signal integrated circuits and systems as well as sensors, with two B.S. courses including the course on Sustainable Development and Transition. He has authored or coauthored more than 150 technical papers and conference contributions and holds three delivered patents. He (co-)received three Best Paper/Poster/Design Awards in IEEE conferences (ICCD 2008, SOI Conf. 2008, FTFC 2014). He serves as a reviewer for various journals and conferences such as IEEE JOURNALS OF SOLID-STATE CIRCUITS, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, IEEE TRANSACTION ON CIRCUITS AND SYSTEMS I/II. Since 2008, he presented several invited papers and keynote tutorials in international conferences including a forum presentation at IEEE ISSCC 2018. On the private side, he pioneered the parental leave for male professors in his faculty to spend time connecting to nature with his family.