Analog Weight Buffering Strategy for CNN Chips

G. Liñán-Cembrano, A. Rodríguez-Vázquez, R. Carmona, S. Espejo and R. Domínguez-Castro
Instituto de Microelectrónica de Sevilla. IMSE-CNM-CSIC
Avda. Reina Mercedes s/n 41012 Sevilla (SPAIN)
Tel.:+34955056679, Fax: +34955056686
E-mail: angel@imse.cnm.es

Abstract[1]

Large, gray-scale CNN chips employ analog signals to achieve high-density in the internal distribution of the template parameters. Despite the design strategies adopted at the circuitry employed to implement the weights, accuracy is ultimately limited by the controlling signals. This paper presents a buffering strategy intended to achieve 8-bit equivalent accuracy in the distribution of the internal analog signals, as employed in the chips ACE4k [1], ACE16k [2], and CACE1k [3].

1. Introduction

Spatial uniformity is a key attribute for correct operation of CNN processors. Such uniformity can be degraded due to mismatch among circuits at different array locations and to errors in the signals that program the operation of the cells. These signals are generated at the chip periphery and distributed across the whole array. The problem of distributing electrical signals to a very large array of cells must be carefully analyzed and understood since it might break up the spatial uniformity required in CNN arrays. The target consists of making "equal" two electrical signals at different, widely separated, locations within the array.

The issues to be confronted are different depending upon wether the signals to be made equal are analog (used for carrying parameters) or digital (used for carrying instructions). This paper focuses on the former. Consider that every building block has been designed for 8-bit equivalent accuracy based on the formulation of systematic and random errors [4]. It basically means that you have designed each block by spending the minimum amount of area which ensures this analog accuracy. Consider now the synapses and suppose that the distribution of the weights is not so accurate. Then, what you really obtain is a system which is less accurate than you expected. The operator – multiplier – provides the required precision but the operand – weight – does not. At the very end, your cell density – which has been penalized by designing more accurate multipliers – is not justified by the performances that you get!

This paper is intended to illustrate the problems of signal distribution to large arrays, and to detail the solutions adopted in ACE4k, ACE16k and CACE1k.

2. Driving an Array of Low Impedance Nodes

Let us start our analysis by illustrating the problem of driving an array of low impedance nodes. This choice is not fortuitous since weight signals are provided to the one-transistor synapses in our chips via one of its diffusion terminals. Hence, the analog buffer providing the weights must also provide the synapse current. Therefore, we must consider the voltage drop across the conductive path which connects the output node of the buffer – somewhere in the programming block – to the diffusion terminal of every synapse.

Consider the case of distributing a voltage level V_w to the array. Let us assume the electrical model in Fig. 1. Here, weight voltages are column-wise transmitted. Assume the very worst case in which all the synapses drive the maximum current I. Let R_u be the impedance of the conductive path which connects the same weight terminal in two adjacent cells in a column. Let R_o be the output impedance of the buffer connected to the output of the DAC, let R_c be resistance of the segment which connects the column to the horizontal line which distributes the weight, finally let us define an additional resistor R_E which accounts for the resistance of the segment between two columns of the horizontal bus. Obviously, the larger the resistivity of the path from a cell to the buffer, the larger the error in the transmitted voltage. If we evaluate the voltage drop between V_w and $V_{N,M}$ – $N = N_{row}$; $M = N_{col}$ – as an upper limit for that error it is found that,

$$V_w - V_{N,M} \approx I \cdot [N_{row} \cdot N_{col} \cdot R_o + \frac{N_{col} \cdot (N_{col}-1)}{2} \cdot R_E \cdot N_{row} + + N_{row} \cdot \left(R_c + \frac{N_{row} \cdot (N_{row}-1)}{2} \cdot R_u]\right) \qquad (1)$$

Let us now assign numerical values for the constants in eq. 1. First of all we need the synapse's current consumption that

[1] This work has been partially funded by DICTAM IST-1999-19007, and CICYT TIC1999-0826

0-7803-7761-3/03/$17.00 ©2003 IEEE

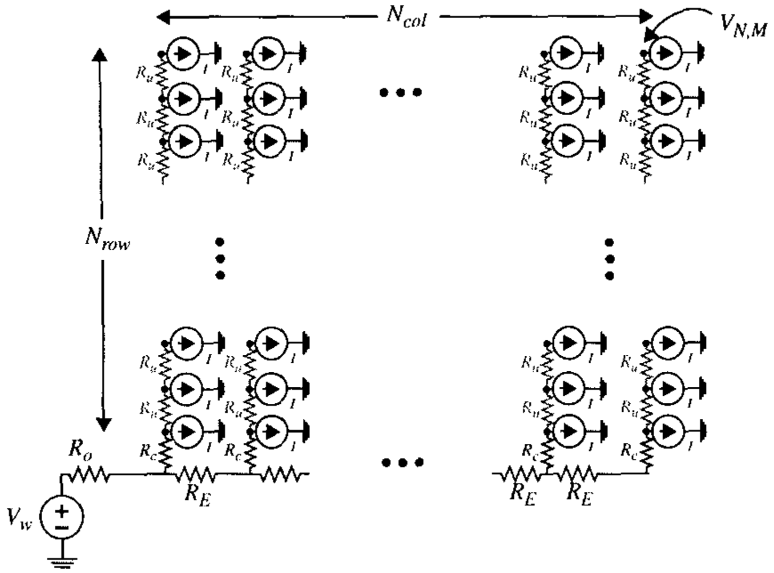

Fig. 1. Model for the Study of the Current Distributing to the Array of Low Impedance Nodes

in the case of the ACE4k prototype is about[1] $1.8\mu\text{A}$. In order to evaluate R_u we can use the following expression

$$R_u = R_{\square m_3} \cdot \frac{L_{cell}}{W_{met}} \qquad (2)$$

where $R_{\square m_3} \cong 30\text{m}\Omega/\text{sq}$ is the sheet resistance of the third metal layer, $L_{cell} = 102.2\ \mu\text{m}$ is the cell height, while W_{met} is the width of the metal layer driving the weight to the cell. Since we need to transmit 20 weights [2] and the width of the cell is $120\mu\text{m}$, the maximum width of each metal line is $6\mu\text{m}$, which after including the separation between lines is reduced to about $5\mu\text{m}$. Using these values in eq. 2 yields,

$$R_u = 0.6\Omega \qquad (3)$$

Without loss of generality, we make $R_c = 1\Omega$ –as it corresponds to assuming that this segment is twice as long as the cell height. On the other hand, since R_o is the output impedance of the buffer, we will use a value of about[3] 0.1Ω. Finally, R_E can be estimated as,

$$R_E = R_{\square m_3} \cdot (W_{cell}/W_E) \qquad (4)$$

where W_{cell} is the cell width and W_E is the width of the horizontal metal line. Assuming that we could use a metal line of $50\ \mu\text{m}$ [4] yields,

$$R_E \cong 60\ \text{m}\Omega. \qquad (5)$$

by using these values in eq. 1, and considering that $N_{row} = N_{col} = 64$ yields,

$$V_w - V_{N,M} = 17\ \text{mV} \qquad (6)$$

Since ACE4k uses a weight signal swing of 800mV, and the desired accuracy[5] is 8-bit the maximum allowable error should be,

$$max(Err) \le \frac{1}{2}\text{LSB} = 1.5625\ \text{mV} \qquad (7)$$

However, the obtained accuracy – even allowing the horizontal bus to occupy 1mm – is only 5.5-bit. This is an important conclusion; it means that, even by using device areas in the cell which guarantee 8-bit accuracy, the obtained precision, because of transmission degradation, is only 5.5-bit. Hence, the increase of area in the analog processing circuitry is not worth. Accuracy is constrained by different mechanisms than mismatch.

[1] This value accounts not only for the signal component but also for the offset term which must be cancelled by the on-cell current memory [1].

[2] This implies to assume that the third metal layer will be fully employed for weight routing purposes.

[3] The values of R_o and R_c do not have important influence in the final result.

[4] Notice that since we use 20 weights, the total width of the horizontal weight bus will be about 1 mm which is a huge value.

[5] We require 8-bit equivalent accuracy for every analog block in the chip.

3. The Distributed Buffer Strategy

Let us now examine eq. 1 to find out a solution to the problem. Easily, one understands that the factor determining the transmission error is the large number of cells connected to the buffer. Indeed, most of the error is due to the horizontal metal line driving the weights to different columns. Here, the influence of the resistance of every segment is multiplied by approximately $N_{col}^2 \times N_{row}$. Hence, a suitable solution consists of minimizing the number of columns connected to the buffer. Such solution relies on the replication of the output branch of the original buffer – properly scaled in order to provide current to a single column – and on connecting such an output branch to every column in the array. Providing those secondary level buffering stages with a high input impedance will avoid for static voltage drops across the lines which drive them. Therefore, by renaming R_o as the output impedance of each of those output stages to find [1],

$$V_w - V_{N,M} = I \cdot [N_{row} \cdot R_o + N_{row} \cdot R_c + \ldots \\ \frac{N_{row} \cdot (N_{row} - 1)}{2} \cdot R_u] \tag{8}$$

where the effect of R_E disappears due to the high input impedance of the secondary stages. The evaluation of this expression yields,

$$V_w - V_{N,M} \approx 3 \text{ mV} \tag{9}$$

which still does not satisfy our requirements.

Next step is to further reduce the number of rows driven by each secondary level buffer. Due to the spatial uniformity required to CNN arrays, the only way to do that is by replicating the secondary level buffers also at the top of each column. Now, the effective number of cells driven by each buffer is halved and we get,

$$V_w - V_{\frac{N}{2},\frac{M}{2}} \approx I \cdot \left[\frac{N_{row}}{2} \cdot R_o + \frac{N_{row}}{2} \cdot R_c + \ldots \\ \frac{\frac{N_{row}}{2} \cdot \left(\frac{N_{row}}{2} - 1\right)}{2} \cdot R_u\right] \tag{10}$$

The evaluation of this new expression, by using $10\ \Omega$ as the output impedance of each elementary column buffer, yields,

$$V_w - V_{\frac{N}{2},\frac{M}{2}} = 1.1 \text{ mV} \tag{11}$$

which satisfies our requirements.

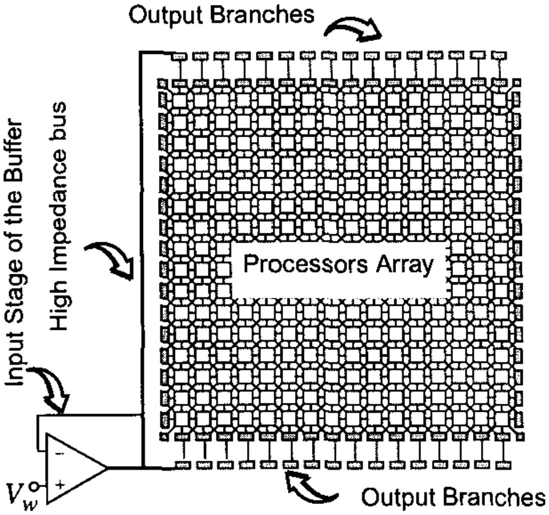

Fig. 2. The new topology of distributed buffers

Fig. 2 shows the block diagram of the distributed buffer topology, while Fig. 3 shows the schematic of the complete buffer. Its first stage consists of a folded-cascode OTA while the output branch is a modified source follower which uses an negative feedback loop in order to lower the output impedance of the buffer as compared to what happens in conventional 2-Transistors source-follower structures. Moreover, the current through transistor M_b does not depend on how much current must be sent to the array since it is fixed by current source M_c. Conversely, in the case of the 2-T source-follower, the current through the transistor which plays the same role as M_b, and which fixes the output impedance of the buffer as $R_o = (gm_b \cdot A)^{-1}$, is given by $I_b = I_{bias} - I_{out}$. Finally, the large capacitance[2] at node Out serves for compensation purposes.

4. The Effect of Mismatch: Horizontal Bus

The distributed buffer topology bases its functionality on avoiding voltage drops across the horizontal buses which

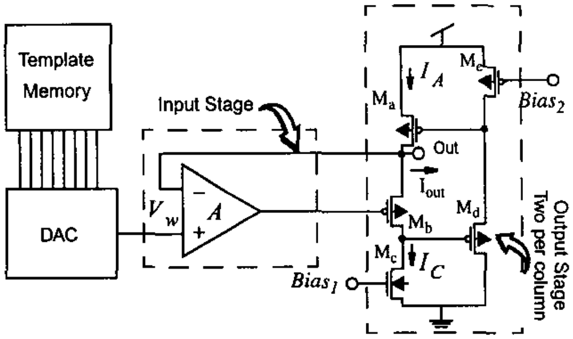

Fig. 3. Distributed Buffer Schematic.

[1] In this case, $V_{N,M}$ is the voltage at the last cell of each column.

[2] $N_{row}/2$ times the input capacitance of the synapses plus the parasitic capacitance of a long line crossing the array.

drive the signals to the output stages located at the top and bottom of every column. However, it relies on a perfect matching between output branches in different columns. Of course, this will be not true in practice.

Let us now consider the circuit in Fig. 4. Here, each output stage is modelled by one transistor – M_b in Fig. 3 – and two current sources I_{Aj}, I_{Cj} – for $j = 1 \ldots N_{col}$ – which account for transistors M_c and M_a in Fig. 3. In addition, we will consider that the current which is required by the *j-th* column is I_{col_j}. If we introduce mismatching effects we can write,

$$I_{Aj} = I_A + \delta I_{Aj} \qquad I_{Cj} = I_C + \delta I_{Cj} \tag{12}$$

where $[\delta I_{Aj}, \delta I_{Cj}]$ are due to random fluctuations of the technological parameters. Ideally,

$$I_{col_j} = I_{Aj} - I_{Cj} \tag{13}$$

however, due to mismatch, each output stage introduces an additional current in the output node given by,

$$\delta I_j = \delta I_{Aj} - \delta I_{Cj} \tag{14}$$

Then, it can be demonstrated that the standard deviation for the difference of the output voltages of two branches is given by,

$$\sigma(\Delta V) = R_E \cdot \sigma I \cdot \sqrt{\sum_{k=1}^{N_{col}} k^2} \tag{15}$$

where [1],

$$\sigma^2 I = \sigma^2 I_A + \sigma^2 I_C \tag{16}$$

and R_E comes to the scene again as the resistance of the segment which connects the outputs of two adjacent output buffers. By using our parameters in eq. 15 we obtain that the required width of the horizontal metal line connecting the output nodes of adjacent output buffers must be [2],

$$W_E > 8\mu\text{m} \tag{17}$$

Therefore, this distributed buffer topology solves two problems at the same time, that of driving large arrays of low impedance nodes, and also helps in reducing the area required for routing lines.

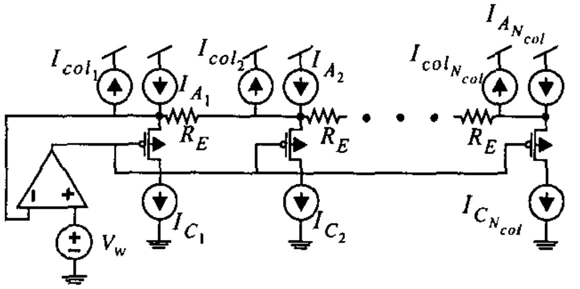

Fig. 4. Schematic for Illustration of the Mismatch Effect

5. Conclusions

This paper has addressed the problem of accurately distributing analog signals to very large arrays of low-impedance nodes. In these systems, non-null metal resistivity leads to voltage drops across the lines steering the signals through the network. This is particularly important for weights, as their spatial uniformity is one of the most characteristics properties of the entire systems. The presented technique, successfully applied to three new generation CNN chips, consists of avoiding the driving high currents through long resistive paths. Instead, high impedance lines and feedback mechanisms are employed in order to short the distance DC currents must travel until reaching the synapses weights terminals. The final topology reduces this distance to half a column of cells – the shortest path which does not require modifying the spatial uniformity of the array – and produces final DC voltage drops of less than 2mV, even if the place where the signal is generated and the place where it is applied are separated by 0.7cm.

6. References

[1] G. Liñán, S. Espejo, R. Domínguez-Castro and A. Rodríguez-Vázquez, "ACE4k: An Analog I/O 64x64 Visual Microprocessor Chip with 7-bit Analog Accuracy". *International Journal of Circuit Theory and Applications*, Vol. 30, pp. 89-116, March-June 2002.

[2] G. Liñán, S. Espejo, R. Domínguez-Castro and A. Rodríguez-Vázquez, "Architectural and Basic Circuit Considerations for a Flexible 128 x 128 Mixed-Signal SIMD Vision Chip". *Analog Integrated Circuits and Signal Processing,* Vol.33, No. 2, pp. 179-190, November 2002.

[3] R. Carmona, F. Jiménez-Garrido, R. Domínguez-Castro, S. Espejo, T. Roska, C. Reckezki and A. Rodríguez-Vázquez, "A Bio-Inspired 2-Layer Mixed-Signal Mixed-Signal Flexible Programmable Chip for Early Vision". *IEEE Transactions on Neural Networks,* (submitted).

[4] A. Rodríguez-Vázquez, G. Liñán, S. Espejo and R. Domínguez-Castro, "Mismatch-Induced Trade-offs and Scalability of Analog Preprocessing Visual Microprocessor". *Analog Integrated Circuits and Signal Processing,* to appear in 2003.

[1] The values for $\sigma^2 I_A$ and $\sigma^2 I_C$ were obtained from Montecarlo simulations once the buffer was already designed.

[2] In the final layout of the chip we employed a $17\mu\text{m}$ wide metal line in order to further reduce this voltage drop.