# DIANA analog core modeling

This project aims to develop a black-box model as a look-up table for the analog core of the DIANA chip. The model can be used to train the neural network or understand the SoC limitations.

Unit time [50, 140, 10], activation value [-63, 63, 1], and VCSbias [0.61, 0.73, 0.5] are the inputs to the model, and the output is the column readout value. First, a plain model of DIANA’s normal behavior will be delivered. In this model, according to exploration experiments, a look-up table is proposed in which having the input values, the output is determined. This naive model is called the zero-order model in this document. In experiments, an n-row section of the analog array is provided by inputs. The inputs are swept across their valid values, and the average value of 512 columns is stored as the output value. Thus, although the output values are integers [-63,63], the model’s output can be a float number. This is in a probabilistic manner, 20.3 is more likely to be 20, but it may also be 21. In the zero-order model, the differences between the readouts of the columns are due to noise and on-chip variation; modeling the former is not possible, and the latter is different from sample to sample, so considering it is not feasible for a generic model of DIANA.

The n-row sections for each VCSbias are selected in a way to have the maximum output swing for values of unit time and activation without output saturation. A higher dynamic range helps for higher resolution in the model, considering that the accumulation in a column is relatively linear (Figure 1). Also, more rows increase the model’s accuracy by lowering the effect of noises and variations.

Chart, line chart

Description automatically generated

Figure 1 Linear accumulation in columns. The number of inputs with constant value increases causes a relative increase in the output.

## General Considerations:

This section presents general results that are assumed to be new. Noticing them may help the user to utilize DIANA more efficiently.

### Bi-Component Model

As Figure 1 shows, the multiply and accumulation operations are fairly linear in the summation lines. However, the sources of error are in 1- spatial variations, i.e. different output with the same input applied on different parts of the array, 2- design imperfections, e.g. IR drop in VCSbias leads to lower sensitivity in the higher columns, and charge bar or ADC error. This error is a function of the output rather than the input or row index. If the error from a fitted line for different experiments is calculated and plotted over the output value (Figure 2), it can be seen that the error is a function of the output. The higher errors for the high output values are due to big VCSbias voltages that will be discussed later.

Chart, line chart, histogram

Description automatically generated

Figure 2 Error is a function of output

The output-related error suggests a bi-component model consists of a linear multiply/accumulation operator and a nonlinear system afterward. This way of modeling is beneficial because the circuit is used with any activation values and numbers as a linear model, and then, a nonlinear function is applied to the output. In other words, a bi-component model reduces the complexity of the nonlinear modeling to a practical level in which, instead of having nonlinearities in all parts of the circuit, it is applied at the output with a restricted range (-31, 31). Figure 3 depicts the bi-component model.



Figure 3 bi-component model

### IR Drop

For many active rows, the output in the columns with higher indices drops in comparison to the columns with lower indices but with the same activations and weights, and so, the same expected output. This is assumed to be due to resistive VCSbias voltage drop across the chip. However, when the loading increases, the output from the column with index zero to higher indices decreases, but after around the column index 300, the trend is reversed, and the output retrieves a part of the drops to the last column. Figure 4 illustrates the variation of drop and retrieve for different number of active rows. For example, it shows that when 800 rows are activated, the output of column 300 is 12 less than column 0 with the exact same expected output. Then, the output of column 512 is 4 more than column 300. The output of different columns with the same input is shown in Figure 5.

Chart, line chart

Description automatically generated

Figure 4 drop and retrieve in higher indices columns

Chart

Description automatically generated

Figure 5 Illustration of column output variation

### Bistable-like behavior

Four experiments are done on 20 consecutive rows in different locations on the array. It seems there are two possible values for the output. In each experiment, the output values can be from one of these two sets (Figure 6). The experiments are conducted with all possible UT values and VCSbias=0.61v.

Chart, line chart

Description automatically generated

Figure 6: Four experiments show there are two sets of possible output values for each input

### Unit Time

At the top rows, a decrease in unit time does not prevent the circuit from going to saturation when the saturation is due to a few high-value input activations. It only reduces both resolution and linearity (Figure 8).

However, as the accumulation is linear, lower unit times can be utilized to increase the output range while deteriorating linearity and resolution. For example, Figure 7 is the same experiment as Figure 8 a) for the double number of activation rows. The optimum unit time index to have a good output swing dropped from 6 or 7 to 3.

Chart

Description automatically generated

Figure 7: input/ output diagram for different values of unit time



Figure 8: a) DIANA with a smaller value for unit time gets saturated almost at the same input as unit time index=9. However, it has less resolution and b) and c) worse linearity.