A Custom FPGA for the Simulation of Gene Regulatory Networks *

Ilias Tagkopoulos, Charles Zukowski
Department of Electrical Engineering
500 W. 120th Street New York
NY 10027, Columbia University
{it2003, caz}@columbia.edu

German Cavelier, Dimitris Anastassiou †
Department of Electrical Engineering
500 W. 120th Street New York
NY 10027, Columbia University
{anastas, cavelier}@ee.columbia.edu

ABSTRACT

We present a unique FPGA that uses a mix of digital and large-signal analog computation for the simulation of gene regulatory networks. The prototype IC consists of a 4x5 array of configurable logic blocks along with programmable interconnect. It can simulate a network of pathways involving up to 20 genes and their associated proteins. The circuit design takes advantage of a number of analogies between CMOS circuits and gene networks. For example, a capacitor charge is used to represent a protein concentration, currents represent protein production, and a transistor switch network is used to compute the influences of activator and repressor proteins on gene expression. A simulation shows how the chip can predict oscillatory behavior in a particular well-known three-gene system.

Categories and Subject Descriptors

B.7.0 [**Hardware**]: Integrated Circuits—*General*

General Terms

Design,Experimentation

Keywords

Gene regulatory networks, custom mixed signal FPGA, genetic pathways

1. INTRODUCTION

Every living cell is a system of enormous complexity, in which numerous molecules interact with each other in multiple ways. Their behavior depends on some external inputs from the environment, as well as on the initial state of the interacting molecules. These interactions are often depicted by graphs defining particular pathways. The result is a precisely coordinated and timed execution of a sequence of cellular processes. It is crucial to understand the particular function of the involved mechanisms and be able to (re)produce the resulting behavior. However, the complexity of these pathways can make flexible simulation on general-purpose machines very difficult.

The approach that we are investigating to allow the simulation of very complex gene regulatory networks is to use a unique analog/digital FPGA that is specifically designed for this purpose. A standard shift-register is used to store configuration information, specifying a model for each gene network being considered. Basic cells are configured to model chemical reactions, such as gene transcription, using a choice of models ranging from simple Boolean logic to classes of nonlinear differential equations. Interconnect is configured to model the influence pathways among various proteins and genes. Analogies between the behavior of gene networks and CMOS circuits are used to make the computation very efficient when compared to what can be done on a general-purpose computer.

2. GENE NETWORK MODELS

The gene regulatory networks that we are simulating are based on the "central dogma" of biology, which states that genetic information flows from DNA to RNA(transcription) and from RNA to protein(translation). Some of these proteins (called transcription factors) are capable of causing the activation or deactivation of a gene when they are bound on specific binding sites on the DNA sequence. Some models that have been proposed to describe the process of transcription and translation use Boolean networks [1], where the expression level is either one or zero in discrete time, or systems of differential equations [2]. The generic set of equations for the latter case is given by (1) and (2)

$$\frac{d[p_i]}{dt} = L \cdot [m_i] - U \cdot [p_i] \quad (1)$$

$$\frac{d[m_i]}{dt} = f(p_1, ..., p_i) - V \cdot [m_i] \quad (2)$$

in which $[m_i]$ and $[p_i]$ are the time-varying mRNA and protein concentrations, *f()* is the non-linear transcription function of i transcription factors, L is a transcription constant, and V and U are degradation rates of mRNA and proteins. The above elements take the form of n-dimensional

*This work was supported by the Department of Energy Grant DE-FG02-01ER25500
†Columbia Center for Computational Biology and Bioinformatics (C2B2)

GLSVLSI'03, April 28–29, 2003, Washington, DC, USA.

matrixes in the case of an n gene system. A particular widely used choice[3], which our design approximates, is to use the Hill function, where *f()* is given by $\frac{a}{1+{p_i}^n} + a_0$.

All the models described above have been implemented in our design, whose basic processing unit can handle four analog input variables and has one analog and three digital outputs. The digital outputs are used for building clusters of processing units which are capable of handling more than four inputs.

3. DESIGN OF THE PROGRAMMABLE SIMULATION MACHINE

3.1 FPGA IC Architecture

Our prototype IC consists of an array of 5x4 custom designed configurable logic blocks (CLB), which are strongly interconnected through switch boxes. A shift register (using TSPL latches) holds information about various weights, model types and connections. Production of proteins and mRNA are represented by currents in an analog fashion, so the buildup of concentrations can be computed by using output currents to charge MiM capacitors at the output of the CLB. The resulting voltages, which represent the concentrations of protein/mRNA are propagated through the interconnect. Since interconnect wires are distributed RC lines, they can naturally model molecule diffusion from one site to another.

3.2 Configurable Logic Block

The CLB is the basic processing unit of our IC. It consists of analog and digital modules and its output is of the form:

$$\frac{d[V_{out}]}{dt} = \sum_{k=1}^{n} \mathbf{x_k} \cdot f(V_{TF_k}) + g(V_{TF_i}, V_{TF_j}) - \mathbf{x_{deg}} \cdot [V_{out}] \quad (3)$$

Where x_k are programmable weights (parameters of discrete values) and V_{TF_k} is the voltage of the k^{th} input. For each CLB , n can be set within the range of one to four, as can i, j, i.e. V_{TF_i}, V_{TF_j} is the voltage of the i^{th} and j^{th} input respectively. Functions $f(V_{TF_k})$ and $g(V_{TF_i}, V_{TF_j})$ are nonlinear and are described in equations (4) and (5):

$$f(V_{TF_k}) = A \cdot (V_{A_k} + B)^2 \quad (4)$$

$$g(V_{TF_i}, V_{TF_j}) = C \cdot (D \cdot V_{TF_i} \cdot V_{TF_j} + E \cdot V_{TF_i} + G \cdot V_{TF_j})^2 \quad (5)$$

V_{A_k} is a variable depending on V_{TF_k} and is given by :

$$V_{A_k} = \begin{cases} -H \cdot V_{TF_k} + K & \text{for } V_{TF_k} < Val \\ V_{TF_k} + \sqrt{L \cdot V_{TF_k}^2 + M \cdot V_{TF_k} + N} & \text{for } V_{TF_k} > Val \end{cases} \quad (6)$$

where A to N are constants, and Val is also a constant which has been calculated to have a value of 1.145 volts for the technology we are using(TSMC 0.25 μm). Equations (3) through (6) give the dependance between the inputs and output of a cell.

Figure 1 is an abstract schematic of the CLB output circuit. The total output current is the sum of the current coming from the analog portion of the CLB (which represents the function $g(V_{TF_i})$ in equation (3) and the current coming from dependent current sources (which represents $f(V_{TF_k})$ in the same equation). The programmable module connected between the output and ground, is used to model the decay/degradation(represented by x_{deg} in equation (3)).

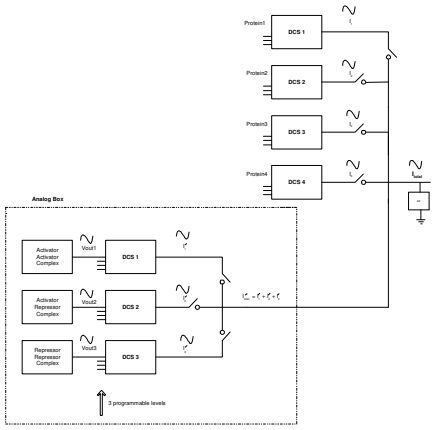

Figure 1: At the output of CLB the currents add, and drive a capacitor. The resulting voltage represents the output concentration of the produced mRNA/protein.

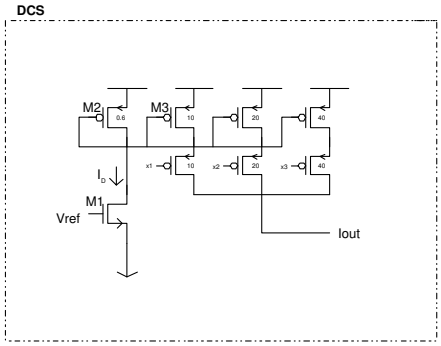

Figure 2: Schematic of a Dependant Current Source

The dependent current source (DCS) works as a current mirror of different weights(one, two, four and combinations of these numbers). As shown in figure 2, an increase of the input voltage V_{REF} would result to a current increase of transistor M_1, due to the quadratic or linear dependance between gate voltage and current. In both regions, when the reference voltage V_{REF} increases, the current I_D of M_1 will also increase. This would result in a drop of the voltage V_A of node (A) because its discharge path has been more strongly turned on. Consequently, M_2 (which will always be in saturation) would turn on further. Transistors M_2 and M_3 form a current mirror and, assuming that M3 will stay in saturation, we would roughly expect that their current relationship will depend solely on their $\frac{W}{L}$ ratios.

3.2.1 *Digital Portion of the CLB*

The digital part of the CLB is shown in figure 3. It consists of four A/D converters and multiplexers. Each A/D converter has two programmable thresholds/outputs, which

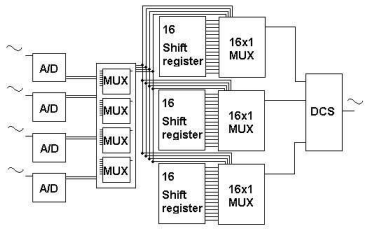

Figure 3: Digital part of the CLB. It consists of A/D conversion, multiplexing, selection and D/A conversion.

are adjustable through 12 control bits from the shift registers. Thus, a total of 8 digitized signals pass through four 8x1 multiplexers and the resulting four signals act as controlling signals at three 16x1 multiplexers, the outputs of which feed a dependent current source (DCS) for a voltage to current conversion. This output is added to the output of the analog part which was shown in figure 1. Thus, depending on which switches(paths) are enabled by the configuration bits, the digital portion of a CLB can either select the weights of a DCS, or function as a boolean network model implementation.

3.2.2 Analog Portion of the CLB

The analog part is responsible for the analog interaction of the various Transcription Factors. It consists of the Activator - Activator, Activator - Repressor, and Repressor - Repressor complexes. The Activator - Activator , Repressor - Repressor subcircuits model the case where two transcription factors are needed to activate/inhibit the transcription of a specific gene, either by binding in separate binding sites, or by forming a dimer. The Activator - Repressor subcircuit is a basic differential pair and can operate in various modes, depending on whether the transcription factors bind on a single or multiple binding sites.

3.3 Interconnect and Timing

The CLBs are amply interconnected through 5x3 switch boxes. Each switch box is an array of 8x12 nodes, each with six transmission gate switches. The overhead of the interconnection network is quite high but adds to flexibility (more about interconnect can be found in [4]). The switch boxes are programmable, controlled by shift registers, and capable of routing signals from the point where they are produced (output of each CLB), to the place where they are needed (input of another CLB). CLBs can be connected to form complex feedback loops. To achieve additional flexibility, between each CLB and a switch box there is a connection box, which serves as a multiplexer.

The interconnect wires and the switch boxes have been carefully designed so that we will be able to simulate the diffusion of the mRNA and protein molecules throughout their trip from the nucleus to the cytoplasm and vice versa. To achieve this goal , we take advantage of the electron diffusion in the wires and inside the interconnect circuitry. The general diffusion equation

$$\frac{\vartheta^2 b(z,t)}{\vartheta z^2} = D \cdot \frac{\vartheta b(z,t)}{\vartheta t} \tag{7}$$

where D is the diffusion constant, describes both voltages in the interconnect and concentrations of molecules within a cell.

It is crucial that the signals driving the analog inputs of a CLB have correct relative timing. Although the delay inside the CLB is fixed, a variable interconnection delay exists and it is caused by the parasitic capacitance/resistance in interconnection wires, switch/connect boxes(transmission gates). To compensate this and to keep the same relative arrival time difference between signals we added programmable capacitances at the input of every CLB. An alternative solution is changing the signal travelling path.

Assuming that the wiring delay inside the switch box and the delay in the connect boxes is insignificant, analysis shows that the total delay of an interconnection with (k) switch boxes and (l) equal line wires would be :

$$\tau_{totaldelay} \approx (l) \cdot \frac{R_{LINE}C_{LINE}}{2} + (k) \cdot C_{eq}R_{eq} \cdot \frac{(N+1)N}{2} + \\ + (l \cdot R_{LINE} + k \cdot R_{eq})C_{prog} \tag{8}$$

where R_{LINE}, C_{LINE} is the total resistance and capacitance of the line, C_{prog} is the programmable capacitance and R_{eq}, C_{eq} the equivalent resistance and capacitance of a transmission gate, which could be assumed constant for all output voltages.

3.4 Design Overview

To summarize , each CLB of our IC is capable of modelling and simulating transcription, translation and several other phenomena. For example it is possible to simulate:

- the mRNA transcription rate as a nonlinear function of a maximum of four transcription factor inputs. Each TF can have different significance.
- the degradation of mRNA and other proteins in the cytoplasm, even if this depends on a specific protein.

Given that many CLBs could ultimately fit on a chip and many chips could cooperate on a board under the supervision of a general purpose computer, our hardware could simulate or a very complex set of pathways and predict its behavior. The top level schematic is shown in figure (4) and the top level layout is shown in figure (5).

4. OPERATION OF THE PROGRAMMABLE SIMULATION MACHINE

Every path and programmable weight in each CLB, has to be set up before running the chip. This means that there are two phases:

- a phase of initialization where we set up all the configurable parameters through the shift registers and we load the protein data.

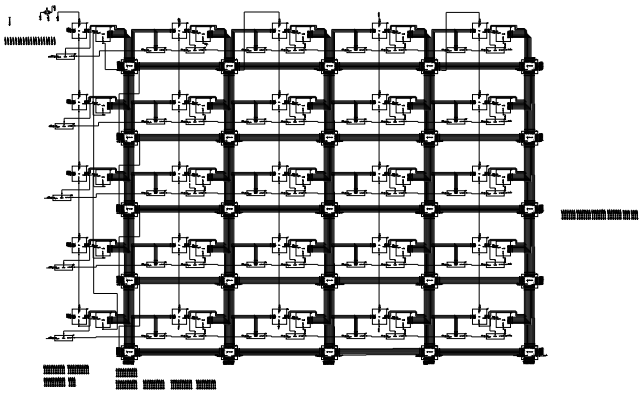

Figure 4: Top level schematic

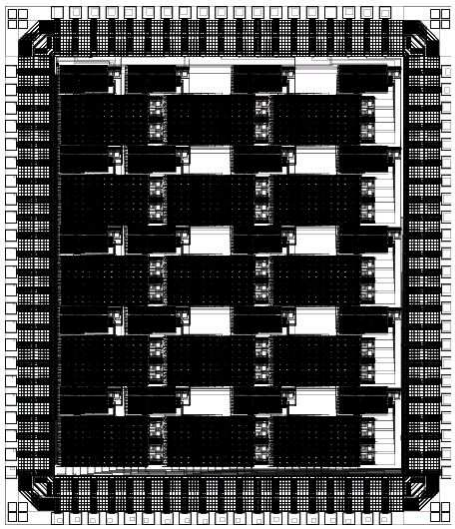

Figure 5: Top level layout in an 84 pin package, TSMC 0.25 μm CMOS mixed signal technology.

- a phase of evaluation, where the initial protein concentrations and the programming of the array will produce a set of protein/mRNA output data.

Each CLB output is connected to a pad and so we are able to measure its voltage. Thus, having 20 CLB's (an array of 5x4) we have 20 measurable voltages/protein concentrations. Moreover, 44 pads have been connected directly to the switch boxes , which enable us to measure the diffusion, delay or intermediate products of a testing pattern.

With our hardware we can simulate a particular set of pathways and predict their behavior. A good and complex example is the repressilator [3], a fictional pathway of three genes, which results in oscillatory behavior of the system. Figure 6 shows the three CLB outputs after running a full software simulation of the chip.

Another application of our hardware would be to run multiple simulations of a system with varying parameters to determine the best model, where simulated outputs would be matched with measured ones. In this case, computation speed is particularly important due to the large number of simulations that might be required.

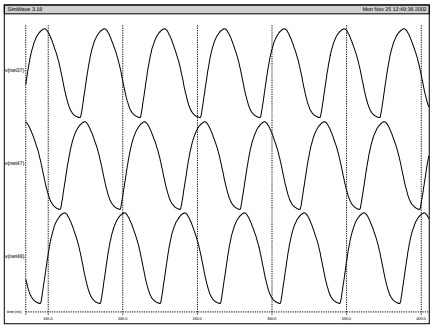

Figure 6: Oscillatory behavior observed in simulation of "repressilator" pattern

5. CONCLUSIONS

In this paper we have presented a custom, mixed digital and large-signal analog FPGA ,which has been designed especially for simulating gene regulatory networks. We have shown that there are significant similarities between the biomolecular and the integrated circuit world, and we have proposed a promising model implementation and circuit for fast simulation. Circuit simulations of the hardware already show that this implementation, although only a prototype, will be capable of simulating complex pathways.

Our ultimate goal is to have an IC which can be configured to simulate a very accurate model of a large biological system. Therefore, in the second version of the IC, we will add a more accurate modelling of biomolecular phenomena. To do that, we have to expand our current work into biochemical networks. Moreover we have also started working on refining the nonlinear transcription function by implementing in silico a physical model of these pathways based on the free energy of the interacting molecules. Finally, we are developing the software needed to automatically generate the configuration bits (shift-register contents) from higher-level network descriptions.

6. REFERENCES

[1] *Akutsu, T., Miyano, S., Kuhara, S.*, Identification of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Net-work Model, Proc. of Pacific Symposium on Bio-computing, pp.17-28, 1999.

[2] *Chen, T., He, H.L., Church, G.M.*, Modeling Gene Expression with Differential Equations, Proc. of Pacific Symposium on Biocomputing, pp.29-40, 1999.

[3] *Michael B.Elowitz, Stanislas Leibler* A synthetic oscillatory network of transcriptional regulators, Nature, 403:335-338, 2000

[4] *Lai,Y.,et al.*Hierarchical interconnection structures for field programmable gate arrays. IEEE transactions on Very Large Integration Systems, vol.5, no.2, June 1997, 186-196.