An Integrate and Fire Spiking Neuron Using Silicon Nano-Wire Technology

Ahmet Bindal and Sotoudeh Hamedi-Hagh

San Jose State University, One Washington Square, San Jose CA 95192, ahmet.bindal@sjsu.edu

ABSTRACT

This study presents a nanometer-scale Integrate and Fire Spike (IFS) neuron cell using vertically-grown, undoped silicon nano-wire transistors. The design cycle starts with determining individual metal gate work functions for each NMOS and PMOS transistor to produce a 300mV threshold voltage. Wire radius and effective channel length are varied to find a common body geometry that yields smaller than 1pA OFF current and produces maximum ON currents for both transistors. Once the optimal device dimensions are defined, a spike neuron cell is built; its transient performance, power dissipation and layout area are measured. Post-layout simulation results indicate that worst-case power dissipation of the neuron is 1.44μW if a single synapse is connected at its output and increases by 18nW per synapse at 500MHz. The neuron circuit occupies approximately 0.116μm^2.

Keywords: vertical FET, neuron, silicon nano wire

1 INTRODUCTION

The early Pulse Rate Coding models were implemented in analog [1] and digital domains [2] using back-propagation algorithm. Recently, a more realistic approach to model neurons is Spiking Response Model (SRM) [3] that directly mimics Hopfield's equation [4]. Most VLSI implementations still favor Leaky Integrate-and-Fire SRM for computational accuracy [5, 6]. Others modified SRM with fuzzy logic to create a new generation of neurons for character recognition [7]. SRM was also reduced to a multi-nanodot floating gate MOSFET form to achieve high density neuron cells in one chip [8].

The primary objective of this study is to design a new ultra-compact Integrate-and-Fire Spiking (IFS) neuron that dissipates substantially lower power and occupies less die area when compared with earlier silicon designs [6] using vertical, undoped silicon nano-wire transistors.

The foundations of silicon nano-wire technology have already been established in terms of material properties and circuit characteristics. CVD-grown silicon nano-wires can be used to fabricate vertical, Surrounding Gate Field Effect Transistors (SGFET) [9]. SGFETs fabricated by conventional processing techniques have also been used in SRAM [10] and high-speed logic circuits [11].

2 NMOSFET AND PMOSFET DESIGNS

Both NMOS and PMOS transistors are designed as enhancement-type with uniform, undoped silicon bodies constructed perpendicular to the substrate. Source/Drain (S/D) contacts are assumed to have Gaussian profiles with a peak doping concentration of 10^{19}cm^{-3}. Both n and p-channel transistors have metal gates and 1.5nm thick gate oxide.

Device simulations are performed using Silvaco's 3-dimensional ATLAS device simulation environment with a 1V power supply voltage. Half of the device is constructed in a two-dimensional platform and then rotated around the y-axis to create a three-dimensional cylindrical structure for simulations. The device radius is changed from 1nm to 25nm while its effective channel length is varied between 5nm and 250nm.

Even though sub-100nm device geometry requires inclusion of Schrödinger's equation to predict electron and hole mobilities, ATLAS device simulator is limited the full usage of such quantum mechanical effects. Instead, this study follows a semi-classical approach in which the semiconductor surface potential is corrected using density gradient method [12, 13].

Vertical and horizontal electric field-dependent mobility models are used to estimate low and high field effects on the current transport. Shockley-Read-Hall recombination and surface recombination models are included to estimate the recombination rates at the bulk and at silicon/oxide interface. Impact ionization model constitutes the only generation model.

The first task of the device design process is to determine an individual metal work function for each minimum length (L_{EFF} = 5nm) NMOS and PMOS transistor to produce a threshold voltage of approximately 300mV. This value constitutes 30% of the power supply voltage and provides sufficient noise immunity for safe large-signal circuit operation. Threshold voltage of the minimum channel length device is first measured as a function of work function for each body radius from 1nm to 25nm as shown in Figure 1. The intersection of threshold voltage with 300mV level in Figure 1 is projected to the x-axis to yield an individual work function value for each NMOS and PMOS transistor at a different body radius.

Both NMOS and PMOS transistors are designed to have leakage currents smaller than 1pA, which is significantly smaller than SOI transistors in earlier modeling studies [14, 15]. Using the work function values for each radius from Figure 1, ON and OFF currents of NMOS and PMOS devices are measured as a function of both device radius and effective channel length. While most transistors within 1nm to 5nm radius range produce I_{OFF} less than 1pA and are considered potential candidates, transistors with larger radii were eliminated because their leakage currents exceeded 1pA as shown in Figure 2.

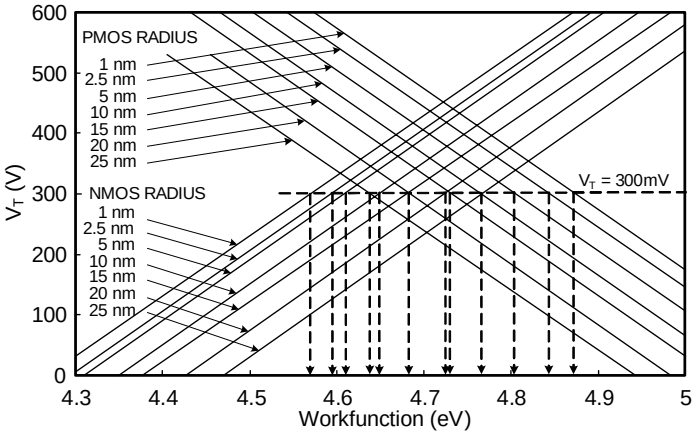

Figure 1: Threshold voltages of nano-wire transistors as a function of metal work function

To reduce the number of candidates in Figure 2, intrinsic transient time of each "qualified" transistor with 1pA OFF current is measured and plotted against ON current in Figure 3. Intrinsic transient time determines the time interval for a fully ON transistor to charge (discharge) the gate capacitance of an identical transistor and is a quick way to understand the transient characteristics of a particular transistor. In Figure 3, ON currents of the qualified NMOS and PMOS transistors start diverging after a device geometry of 4nm radius and 40nm effective channel length; larger wire radius provides higher I_{ON} values for NMOS transistors, but reaches a saturation plateau for PMOS transistors. Therefore, a device geometry of 4nm radius and 40nm effective channel length is considered optimal to produce approximately equal drive currents and reasonably low intrinsic transient times for both NMOS and PMOS transistors.

3 IFS NEURON

Figure 4 shows the firing principle of a single IFS neuron. The neuron cell or "soma" is composed of three parts: the "sum" forms a post-synaptic potential; the "threshold" determines the maximum amplitude for the post-synaptic potential to generate a pulse, the "fire" forms a positive or a negative amplitude pulse according to the amplitude of the post-synaptic potential. Synaptic pulses at the input of a neuron cell can be "excitatory" or "inhibitory" depending on their amplitude. A 1V input pulse defines the excitatory input while -1V defines the inhibitory input. In this figure, four 1000ps wide excitatory synaptic pulses are asynchronously received by dendrites. A post-synaptic potential is formed at the sum node of the neuron; if the peak of the post-synaptic potential exceeds a certain threshold, the neuron fires a 1000ps wide pulse for the other neurons. Similarly, a negative post-synaptic potential may be accumulated; if this potential exceeds a negative threshold, a -1V amplitude, 1000ps wide pulse is fired at the output.

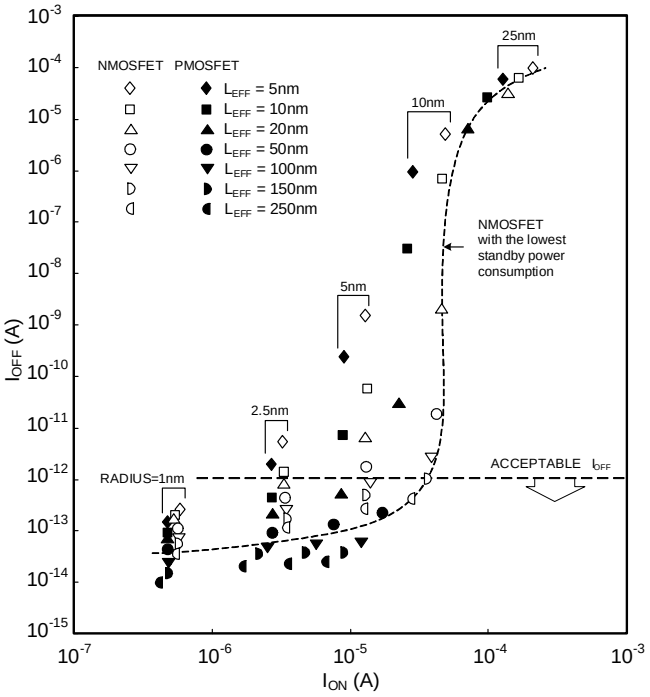

Figure 2: OFF vs. ON currents of nano-wire transistors

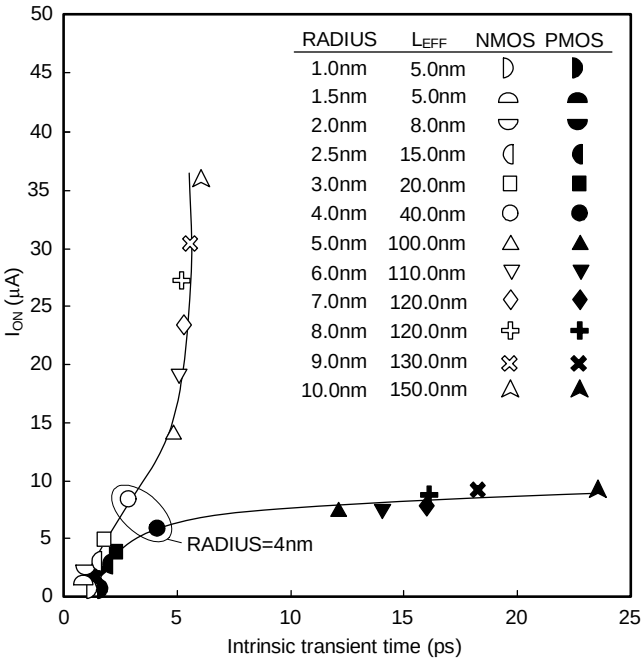

Figure 3: The ON current vs. intrinsic transient time of the "qualified" nano-wire transistors with 1pA leakage current

The IFS neuron schematic is shown in Figure 5. The circuit is composed of four sections where each section is bounded by dashed lines in the schematic.

NSTI-Nanotech 2007, www.nsti.org, ISBN 1420061828 Vol. 1, 2007

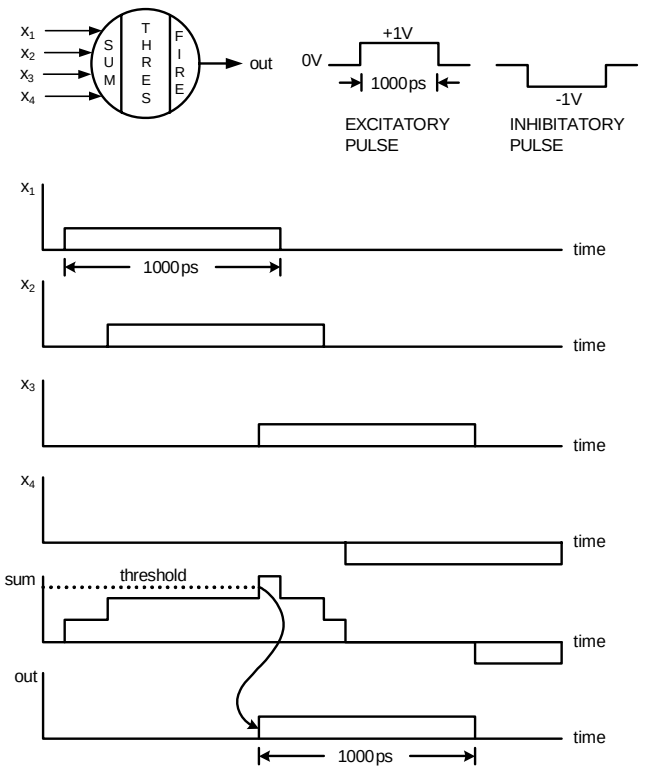

Figure 4: The IFS neuron principle

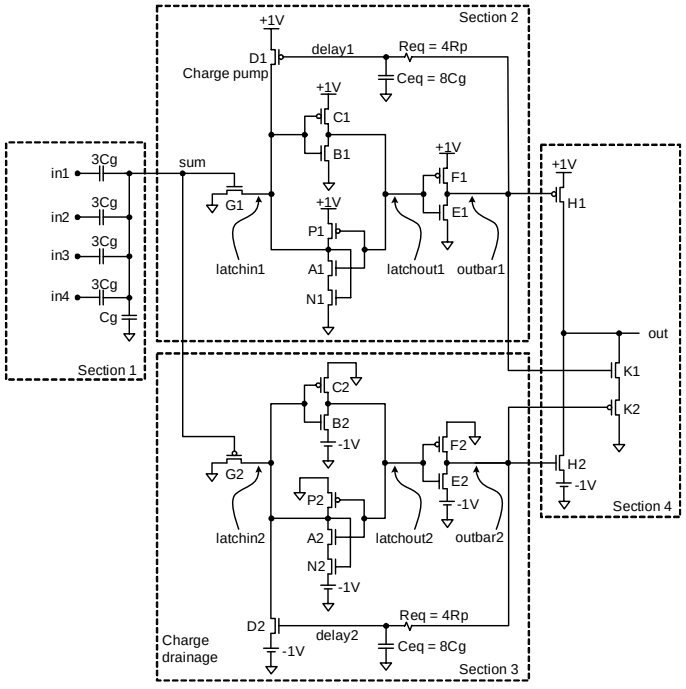

Figure 5: The IFS neuron schematic

The first section is where all synaptic inputs are connected to a common node ("sum" node) via capacitors to form a post-synaptic potential. The cell threshold voltage is determined by capacitive charge-sharing and simply changed by altering the total capacitance value at the "sum" node. For this particular neuron circuit, +/-0.6V is defined as the threshold voltage which generates either 1V or -1V output pulse, respectively. In this figure, Cg corresponds to the gate capacitance of a single NMOS/PMOS transistor, which is equal to 32aF.

The second section is a self-time circuit composed of a latch followed by RC delay element and a simple PMOS charge pump to produce a 1V output pulse. The waveforms generated by this section are shown in Figure 6 with contoured arrows, each of which indicates how the next waveform is produced.

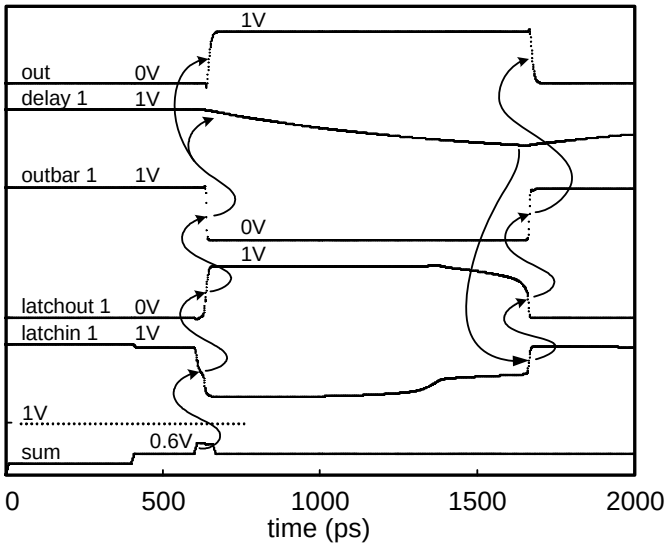

Figure 6: Waveforms of an IFS neuron generating excitatory and inhibitory pulses.

When post-synaptic potential at the "sum" node exceeds 0.6V, G1 turns on and triggers the "latchin1" node from 1V to 0V. Since the "fighting" between G1 and P1 is significantly reduced by the addition of N1, "latchout1" node transitions to 1V, E1 turns on and pulls down the "outbar1" node to 0V after a short delay. Subsequently, H1 turns on and charges the output node to 1V. Following the voltage drop at the "outbar1" node, the voltage at "delay1" node also decays towards 0V with an RC delay specified by Req = 4Rp and Ceq = 8Cg, and eventually turns on D1 to pump charge to the "latchin1" node. This node transitions to 1V which settles in the latch, changes the "outbar1" node back to 1V and turns off H1; the output is pulled down to 0V. The width of the resultant output pulse in Figure 6 is measured approximately 1030ps.

The third section also represents a self-time circuit composed of a latch, an RC delay element and an NMOS charge drainage device, which generate a -1V, 1000ps output pulse when post-synaptic potential at the "sum" node exceeds -0.6V.

The fourth section is the output buffer stage responsible of maintaining the cell output at 0V when both H1 and H2 are turned off. However, when post-synaptic potential at the "sum" node exceeds 0.6V, this stage charges the output node to 1V through H1; when the voltage drops below -0.6V it discharges the output node to -1V through H2.

The output pulse width is adjusted as a function of fixed R and C values in the delay element. A capacitor value is formed only by multiples of Cg = 32aF. Similarly, the resistor is formed either by a series of p-type nano-wires, each of which produces approximately 1.07MΩ or a series of n-type nano-wires, each of which produces 0.75MΩ. In this particular IFS neuron circuit, the pulse width is determined by $T_{PW} = 1.2T_{RC} + 80$, where, T_{RC} is the time constant composed of individual R and C values indicated above.

The pulse generated at the output of IFS neuron has a rise time and fall time according to $T_R = 3.60N + 1.96$ and $T_F = 3.18N + 4.66$ in ps, respectively, where N is the number of neuron synapses connected at the output. The worst-case dynamic power dissipation of the neuron is also represented by N and is equal to $P = 0.018N + 1.423$ in μW.

Figure 7 shows the layout of the IFS neuron with a dimension of 524nm by 221nm. The transistor names are indicated on the layout to match the schematic given in Figure 5. The "sum" node is protected under power and ground lines from any capacitive coupling that may alter the threshold value and cause neuron to fire. A recent 6-transistor SRAM cell designed in a 65nm technology occupied a cell area of $0.57\mu m^2$ [16]. The layout area of the neuron is approximately $0.116\mu m^2$ which is about 5 times smaller than the 6-transistor SRAM cell and contains 9.8 times more transistors.

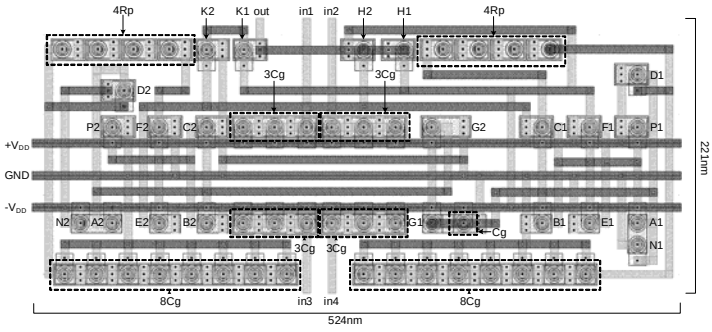

Figure 7: The IFS neuron layout

4 CONCLUSIONS

In this exploratory work, a new IFS neuron is designed using vertical, undoped silicon nano-wire transistors with 40nm channel length and 4nm radius. This optimum device geometry produces 1pA OFF current and generates ON currents in the neighborhood of 10μA. Threshold voltage of each transistor was adjusted to 300mV by fine-tuning the individual gate metal work function. The IFS neuron circuit is operated with excitatory and inhibitory synaptic input pulses whose amplitudes are 1V and -1V, respectively. The neuron cell is capable to produce excitatory and inhibitory pulses at its output between 80ps and 1000ps according an adjustable RC time constant. The worst-case transient time at the output is roughly 8ps for single synapse and increases by 3ps per additional synapse. The worst-case average power dissipation of the IFS neuron is 1.4μW with a single synapse and increases by 18nW per synapse. The neuron cell occupies a layout area of $0.116\mu m^2$.

REFERENCES

[1] T. Morie, Y. Amemiya, IEEE J. Solid State Circuits, Vol. 29, No. 9, p. 1086-1093, 1994.

[2] M. Yasunaga et al., IEEE J. Solid State Circuits, Vol. 28, No. 2, p. 106-113, 1993.

[3] W. Maass, C. M. E. Bishop, Pulsed Neural Networks, Cambridge, MA, MIT Press, 1999.

[4] J. Hertz, A. Krogh, R. G. Palmer, Introduction to the Theory of Neural Computation, Addison Wesley Publishing Co., 1993.

[5] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, D. Amit, Neural Comput., Vol. 12, p. 2227-2258, 2000.

[6] G. Indiveri, E. Chicca, R. Douglas, IEEE Trans. Neural Networks, Vol. 17, No. 1, p. 211-221, 2006.

[7] T. Yamakawa, IEEE Trans. Fuzzy Systems, Vol. 4, No. 4, p. 488-501, 1996.

[8] T. Morie, T. Matsuura, M. Nagata, A. Iwata, IEEE Trans. Nano., Vol. 2, No. 3, p. 158-164, 2003.

[9] B. Yu, L. Ye, M. Meyyappan, NSTI, Vol. 3, p. 232-235, 2005.

[10] T. Kikuchi et al., IEDM, p. 923-926, 2004.

[11] H. Takato et al., IEEE Trans. Elec. Dev. Vol. 38, No. 3, p. 573-578, 1991.

[12] A. Wettstein, A. Schenk, W. Fichtner, IEEE Elec. Dev., Vol. 48, No. 2, p. 279-283, 2001.

[13] A. Asenov, G. Slavcheva, A. R. Brown, J. H. Davies, and S. Saini, IEEE Trans. Elec. Dev, Vol. 48, No. 4, p. 722-729, April 2001.

[14] K. Kim, K. K. Das, R. V. Joshi, C-T Chuang, IEEE Trans. Elec. Dev., Vol. 52, No. 5, p. 980-986, 2005.

[15] J-W Yang, J. Fossum, IEEE Trans. Elec. Dev., Vol. 52, No. 6, p. 1159-1164, 2005.

[16] P. Bai et al., IEDM, p. 657-660, 2004.