An FPGA Architecture Using Vertical Silicon Nanowire Transistors

A. Bindal*, D. Wickramaratne*, S. Hamedi-Hagh* and T. Ogura**

*San Jose State University, San Jose, CA, USA, ahmet.bindal@sjsu.edu
**Halo LSI, Hillsboro, OR, USA, togura@halolsi.com

ABSTRACT

This study presents an FPGA architecture using vertical silicon nanowire transistors. Cylindrical surrounding gate MOS devices of 2nm in radius and 10nm in length are used to produce ultra-low power CMOS circuits for the particular FPGA architecture. Each FPGA cluster consists of three 4-input Look-Up-Tables (4LUT) and a highly reconfigurable bus composed of eight interconnecting wires for cluster-to-cluster connectivity. Post-layout simulation results indicate that the worst-case propagation delays of a 4-LUT are 62ps during read and 68ps during write operations; the worst-case propagation delay of a cluster increases to 72ps for a fan-out of 1 and 97ps for a fan-out of 3 identical clusters. The worst-case power dissipation is approximately 3.1µW for a 4-LUT and 10.2µW for a cluster at 10GHz. The cluster layout which contains three 4-LUTs occupies approximately 8.0µm^2.

Keywords: nanowire, silicon, fpga, architecture, low-power.

1 SGFET DESIGN

Both NMOS and PMOS transistors are enhancement-type with undoped, cylindrical silicon bodies perpendicular to SOI substrate used mainly for latch-up prevention. Each transistor has 2nm thick gate oxide and a metal gate tailored to produce 300mV threshold voltage to provide sufficient noise immunity for a 1V power supply operation. Detailed cross section and layout of a single transistor is shown in Figure 1. 3D device simulations including quantum mechanical effects [1] are performed to obtain NMOS and PMOS I-V characteristics and circuit models.

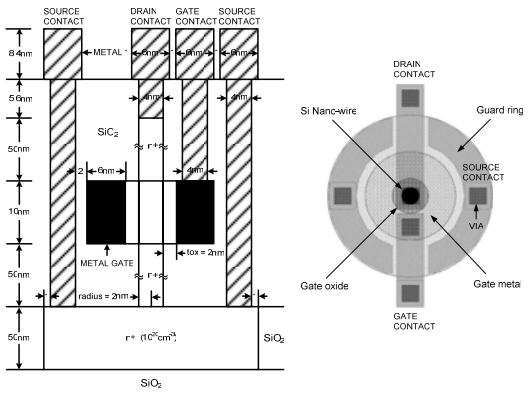

Figure 1: A nanowire transistor cross-section and layout.

The first task of the transistor design is to determine metal work function values for each NMOS and PMOS transistor to produce a threshold voltage of 300mV. This design process is described in detail in [2, 3] for device radii between 2nm and 20nm and shown below in Figure 2.

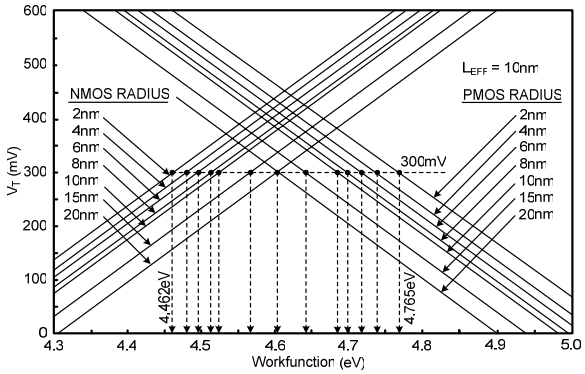

Figure 2: Threshold voltage versus metal work function.

Channel length of each transistor is varied until 1pA or smaller static OFF current is obtained at each radius. Intrinsic transient time and intrinsic energy of each NMOS and PMOS transistor are subsequently measured and plotted against each other to select the best device geometry as shown in Figure 3.

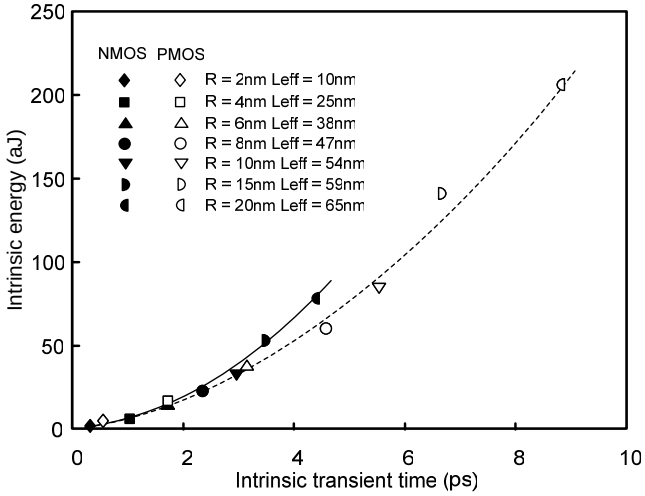

Figure 3: Intrinsic energy versus intrinsic transient time of the transistors with $I_{OFF} \leq 1$pA.

In Figure 3, intrinsic transient time determines the time interval for a transistor to charge/discharge the gate capacitance of an identical transistor when it is fully on and

NSTI-Nanotech 2008, www.nsti.org, ISBN 978-1-4200-8503-7 Vol. 1

is a quick way of measuring the ON current characteristics of a transistor. Intrinsic energy, on the other hand, corresponds to the integration of instantaneous power delivered (received) to (from) the gate capacitance of an identical transistor as a function of time and measures the dynamic power dissipation of a transistor. The most desirable transistor geometry is found to have 2nm radius and 10nm effective channel length configuration which produces minimal gate capacitance as expected.

2 FPGA ARCHITECTURE

The FPGA architecture consists of an array of clusters interconnected in a network as shown in Figure 4.

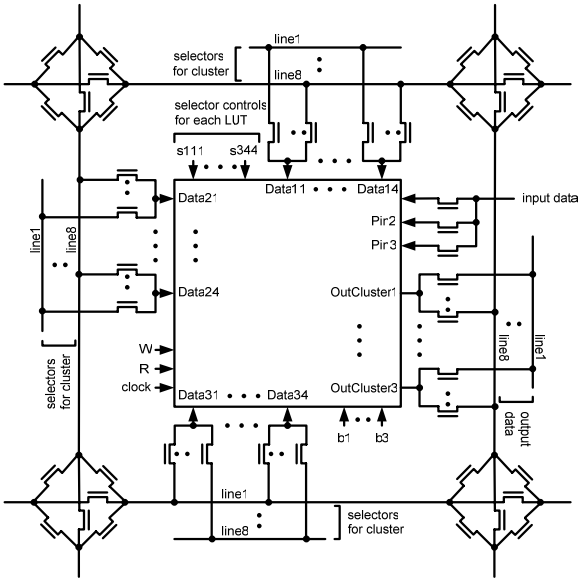

Figure 4: The FPGA architecture.

In this architecture, there are eight interconnecting wires between clusters; each interconnecting wire uses 6-transistor "traffic pole" switch (TPS) at the corners of a cluster to make the overall inter-cluster wiring highly configurable. Each inter-cluster wire is connected to cluster input by an 8-1 pass-gate MUX. Each cluster produces three outputs, each of which is connected to adjacent clusters with interconnecting wires and TPS. Each cluster contains three 4-Look-Up-Tables (4-LUT) as suggested by [4]. Each LUT has four data inputs; its output can be registered, routed to the neighboring LUTs or other clusters via bypass paths as shown in Figure 5. This flexible configuration produces complex gate-level implementations as well as state machines.

The 4-LUT is composed of 16 memory cells, each of which is connected to a single output using an array of pass-gate transistors in the form of a large 16-1 MUX as shown in Figure 6. The same pass-gate configuration also serves as 1-16 DEMUX connecting a single write input to any one of 16 memory cells.

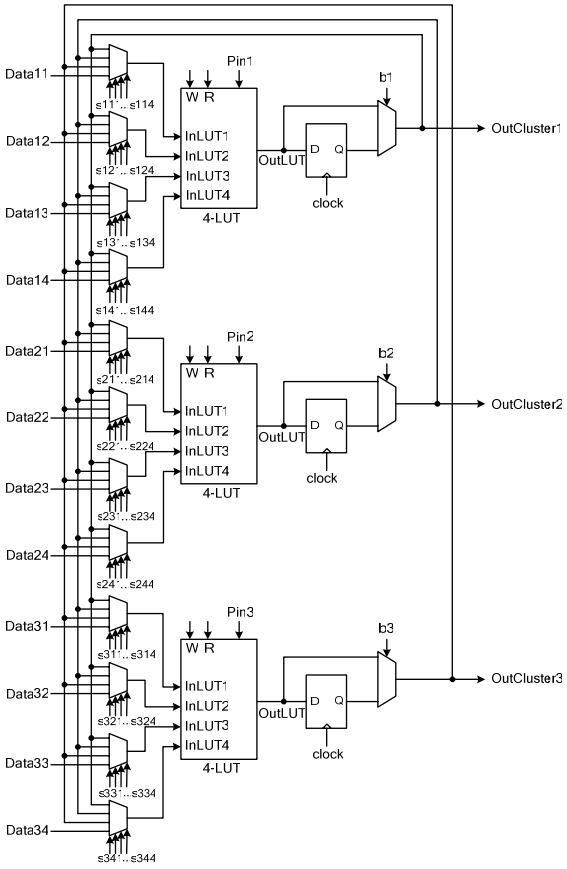

Figure 5: The cluster architecture containing three 4-LUTs.

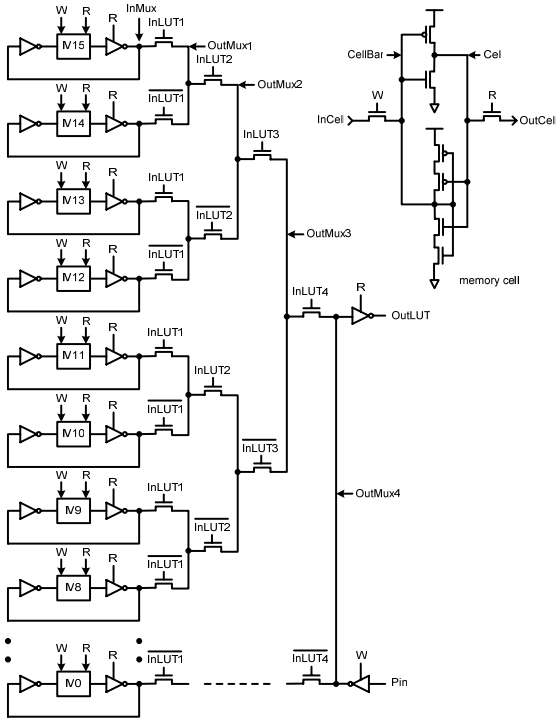

Figure 6: The 4-LUT circuit (M0-M15 are memory elements shown in the inset).

3 CIRCUIT DESIGN

3.1 Circuit operation and waveforms

The waveforms obtained during write operation where logic 1 is written into a memory cell is shown in Figure 7.

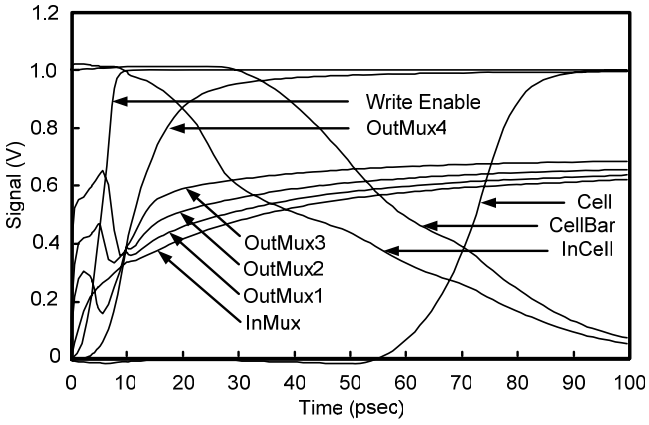

Figure 7: Waveforms of a 4-LUT during write operation.

The node names in Figure 7 are indicated in Figure 6 for cross reference. To write data into a memory cell, the rising signal at the "Pin" input propagates through the write-enabled (W-enabled) tri-state inverter and four pass-gate transistors in series (selected by InLUT1 through InLUT4) prior to arriving the "InMux" node. Figure 7 shows the successive deterioration of this signal by displaying the waveforms at each node from "OutMux4" to "InMux" every time the signal passes through a pass-gate transistor. The waveform at the "InMux" node shows both voltage level and slow rise time problems. The inverter at this node re-establishes the voltage levels of this signal and eliminates the slow rise time before the signal is sent to the memory cell. The re-conditioned signal at the output of the inverter (the "InCell" node) propagates through a pass-gate transistor to the "CellBar" node of the memory cell and is stored at the "Cell" node. In order to reduce the contention between the PMOS transistor of the "re-conditioning" inverter and the NMOS transistor of the "memory cell" inverter (or the NMOS transistor of the "re-conditioning" inverter and the PMOS transistor of the "memory cell" inverter), two extra transistors are added to the cell inverter in the feedback loop as shown in the inset of Figure 6.

The waveforms obtained during read operation where logic 1 is read form a memory cell is shown in Figure 8. The node names in Figure 8 are also indicated in Figure 6 for cross reference. To read data from a memory cell, the tri-state inverter next to the memory cell and the one at the output of 4-LUT must both be read-enabled. Similar to the write operation, the falling signal at the "InMux" node propagates through four pass-gate transistors in series and arrives at the "OutMux4" node. Even though Figure 8 shows some signal degradation in terms of slow fall times at the nodes "OutMux1" through "OutMux4", this deterioration is not as severe as losing part of the signal level due to a threshold voltage drop; NMOS pass-gate transistors do not allow any threshold voltage drop when transmitting logic 0. The slow rising signal at the "OutLUT" node is due to a large capacitive load at this terminal.

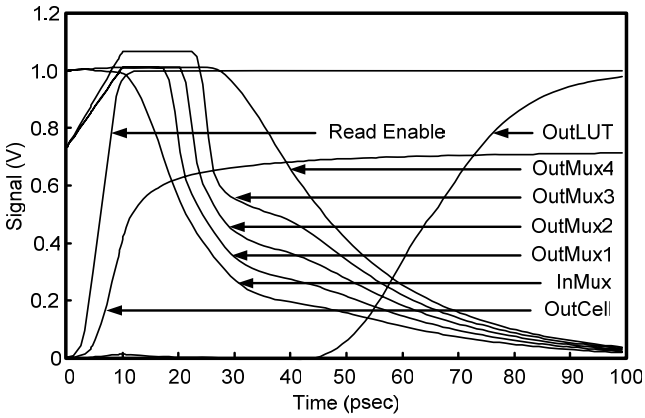

Figure 8: Waveforms of a 4-LUT during read operation.

The inter-cluster bus architecture composed of eight interconnecting wires may also be exposed to signal degradation when transmitting logic 1 through a series of NMOS pass-gate transistors. Connecting two clusters diagonally spaced for more than four cluster lengths is not recommended due to threshold voltage drop in logic 1 level and progressively longer transition times (slow nodes) as the number of pass-gate transistors increase. Figure 9 shows inter-cluster waveforms for transmitting logic 1 from one cluster to the next diagonally placed three cluster lengths away. The stored bit at the cluster of origin shifts from 1V to 0.7V when it goes through the pass-gate transistor of the memory cell and arrives at the "OutCell" node. However, the voltage level is re-established by the R-enabled tri-state inverter at the "InMux" node despite a long fall time. The rising signal at the "OutLUT" propagates through 2-1 MUX placed at the output stage of the cluster as shown in Figure 5 and arrives at the "OutCluster" node. The signal that departs from the "OutCluster" node propagates through seven pass-gate NMOS transistors (one at the 1-8 output DEMUX, five at 6-transistor switch boxes and one at the 8-1 input MUX) and arrives one of the "Data" terminals (Data11 through Data34) of the destination cluster; it suffers from threshold voltage drop and exhibits a very slow rise time as expected. Both of these issues are resolved by 4-1 MUX inverters at the cluster input as shown in Figure 5; progressively better waveforms are generated at the "DataBar" and "InLUT" nodes of the 4-1 MUX before the signal is allowed to

NSTI-Nanotech 2008, www.nsti.org, ISBN 978-1-4200-8503-7 Vol. 1

propagate through the rest of the destination cluster as shown in Figure 9.

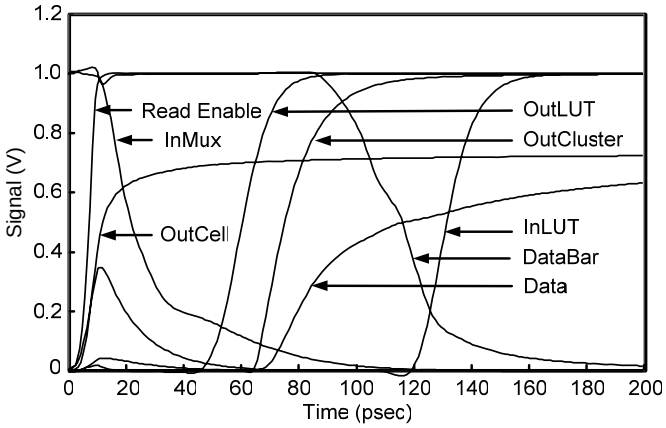

Figure 9: Typical waveforms obtained in a 4-LUT, a cluster and on inter-cluster wires connecting two clusters diagonally three cluster lengths apart.

3.2 Delays and power dissipation

Post-layout, worst-case propagation delays between Read Enable port of a cluster and its output are measured as a function of fan-out (the total number of clusters connected to a cluster output). The rise delay changes with T_R = 9FO + 61 in ps where FO corresponds to fan-out; the fall delay similarly changes with T_F = 12.5FO + 59.5 in ps.

Post-layout, worst-case wire delays are also measured between a cluster output and a neighboring cluster input as a function of fan-out and diagonal inter-cluster distance. The worst-case interconnecting wire delay is T_W = 4FO + 23 in ps for 1 diagonal cluster spacing and T_W = 14.5FO + 54.5 in ps for 4 diagonal cluster spacing.

The worst-case, average dynamic power dissipation is measured at 10GHz for a 4-LUT and a cluster. Power dissipation of a 4-LUT is 2.15μW during write and 3.08μW during read. Similarly, power dissipation of a cluster is 7.09μW during write and 10.16μW during read. Both 4-LUT and cluster power dissipations during read cycle are approximately 40% higher compared to write.

3.3 Cluster layout

The layout of a single cluster is shown in Figure 10. All three 4-LUTs in the cluster are stacked on top of each other; each 4-LUT occupies in the neighborhood of 2.6μm^2 layout area. The total layout area of a cluster is approximately 8.0μm^2.

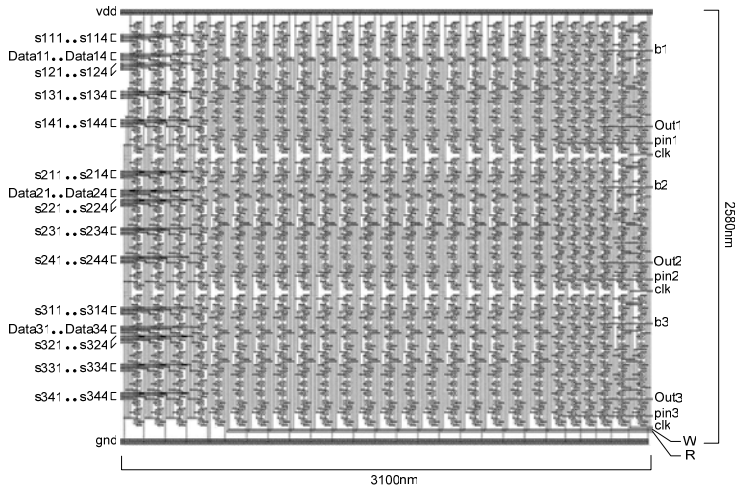

Figure 10: The layout of a FPGA cluster containing three 4-LUTs.

REFERENCES

[1] A. Wettstein, A. Schenk, W. Fichtner, "Quantum Device-Simulation with Density Gradient Model on Unstructured Grids", IEEE Elec. Dev., Vol. 48, No. 2, p. 279-283, 2001.

[2] A. Bindal, S. Hamedi-Hagh, "The Design of a New Spiking Neuron Using Silicon Nano-Wire Technology", Nanotechnology (Institute of Physics), Vol. 18, 095201, 2007.

[3] A. Bindal, A. Naresh, P. Yuan, K. K. Nguyen, S. Hamedi-Hagh, "The Design of Dual Work Function CMOS Transistors and Circuits Using Silicon Nano-Wire Technology", Trans. IEEE Nano., Vol. 6, No. 3, p. 291-302, 2007.

[4] E. Ahmed, J. Rose "The Effect of LUT and Cluster Size on Deep-Submicron FPGA Performance and Density", IEEE Trans. VLSI, Vol. 12, No. 3, p. 288-298, 2004.