Extended Abstracts of the 2004 International Conference on Solid State Devices and Materials, Tokyo, 2004, pp. 108-109

H-1-1 (Invited)

Nanoelectronics and Nanodevices: Issues in Future Information Processing

David K. Ferry

Center for Solid State Electronics Research and Department of Electrical Engineering
Arizona State University, Tempe, AZ 85287-5706, USA

1. Introduction

With the continuing advance of VLSI technology, the size of individual transistors has been reduced significantly. Today, chips already are made with transistors that have nanometer scale, and are a driving force for nanotechnology. This will continue to be the case for many years in the future. Yet, new and novel silicon nanodevices have appeared and others have been suggested. The behavior of such small devices, with characteristic lengths on the 5-20 nm scale is described by quantum mechanics, and this brings new limitations into play. Even so, there are many novel technologies which are claimed to have a future that will supplant Si as the dominant technology. In this talk, we discuss a number of limitations which suggest that this will be difficult to achieve. We will discuss the limitations that arise from power considerations in ultra-dense VLSI, as well as effects which arise in small devices. Some novel devices offer new functionality, and these require a correspongind new architecture, which will be discussed below.

2. Power Limitations

Within less than a decade, the industry will reach the 22 nm node, with gate lengths of 10-11 nm. Beyond this point, it is not at all clear that silicon CMOS, even with high-K dielectrics and including novel materials, will be usable for further scaling. Already, problems with heat dissipation have arisen. In fact, if N is the number of devices per square cm., E is the energy required to switch, f is the frequency of the clock, and P is the probability that a switch occurs in each clock cycle, then we must have $EfNP < W$, the power dissipation per square centimeter that can be tolerated. This is a problem as current scaling trends will push this result up against the thermal (kT) limit at, or near, the 22 nm node. Other limits are kT itself is a limit on energy per switching event and the possible quantum limit in terms of speed. These constraints are shown in Fig. 1, where we also plot the "roadmap" values for comparison. It is clear that we are up against a wall in terms of power and energy. In addition, there is a problem with interconnect lengths, and the dramatic increase in wire length that arises with scaling. The importance of these two issues lies in the fact that *novel devices, which perform only the normal logic operation, will not be able to impact future Si integration*. Rather, we require novel architectures which are optimized for new functional operation of any new, novel devices and which provide a more local type of interconnect—locally-interconnected architectures. We discuss the architecture issue below.

3. Device Limits and Novel Devices

As device size is reduced below the current 90 nm node (45 nm gate length), the problems of discrete impurities and short-range particle-particle interactions become more severe. The device problems when including the short-range interaction between particles is provided by examining the current flow through the device in simulations, as illustrated in Fig. 2. Here, the current density is seen to vary by more than an order of magnitude across the channel as a result of the influence of the particle-particle interactions. As a result, electrons tend to be forced into closer proximity to the dopant ions, which further influences the electron-electron interaction and the energy relaxation behavior. The effect of the inhomogeneity of the dopants becomes apparent by examining the current paths in the drain, which follow regions of higher dopant density. Also, variations in the position of the donor atoms in the source allow those donors that are closest to the channel to serve as "leaders" from which the electron flow in the channel originates.

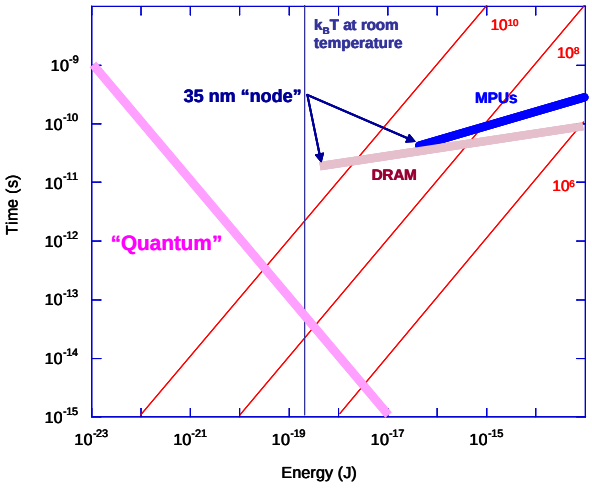

Fig 1 The energy and power limitations are displayed here. The curves sloping down from the upper right are curves of constant density, while that sloping down from upper left is the quantum limit. The vertical line is kT. The roadmap values down to the 35 nm node are shown.

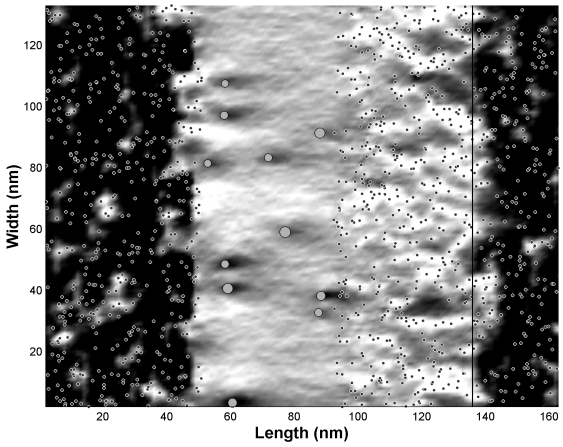

Fig 2 Current flow through a 50 nm gate length SOI MOSFET. The bright areas are regions of higher current density, while the dots and circles are donors and acceptors, respectively [7].

While the above can be a problem, it also opens the door to novel devices, such as single-electron devices,

nanowires, and quantum dots. These may applications as new logic devices or in new areas such as quantum computing. It is fairly well understood that going much beyond the 22 nm node, which will have gate lengths in the 10-11 nm range, is not likely to be possible with our current understanding of CMOS structures, even with SOI and vertical devices (FINFETs, trigates, etc.). As a result, the door is open for these novel devices, especially if they provide new functionality or dramatically reduce the number of devices required for a given function (e.g., move downward in Fig. 1. Some of these devices, including full quantum treatment of the wire-based MOSFET will be discussed. The quantum transport in these ultrasmall structures will lead to new and different behavior, which must be understood before these devices can be fully incorporated into new applications. One such problem is vortex formation, which is depicted in Fig. 1, which leads to circulating currents in the device, and hence to fluctuations in conductance.

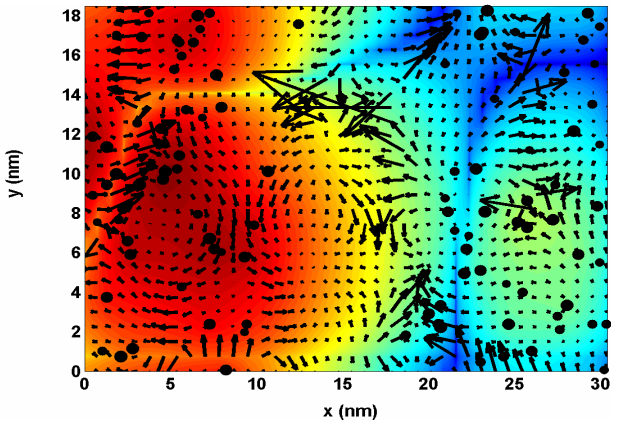

Fig. 3 Vortex formation in a quantum wire SOI MOSFET. The source extends to 11 nm (in the x direction), and the drain begins at 22 nm. Here, a gate voltage of 0.5 V inverts the channel region. Arrows give the velocity vectors [1].

4. Novel Architectures

From the system level limitations shown in Fig. 1 above, it is clear that merely replacing a CMOS gate with an equivalent gate using novel devices will not solve the problem. Rather, new architectures which provide greater functionality per device are required. The number of gates and pins are related to the information flow and a fractal dimension for the chip [2]. The relationship between pins P and gates A is given by Rent's rule, expressed as $P = AG^s$, with a value of $s = 0.3$ [3]. This is related to the fractal dimension through $D = 1/(1 - s) \sim 1.4$, which indicates an information flow less than the two dimensions of the chip [2]. *If we are to make major changes in the on-chip architecture in order to increase functional performance, without significantly increasing the number of devices (through the use of novel devices), then we must adapt to an architecture whose information dimension is closer to 2.* Cellular arrays are such architectures.

The earliest concept of a cellular array was von Neumann's cellular automata [4]. Such a machine could perform computation provided that 2 states, and a sufficiently large neighborhood, were used [5]. This defines two processes: (1) the functional behavior of each cell $f(\bullet)$, which describes how the input maps into an output, and (2) the neighborhood Ω of each cell, which is the group of cell outputs connected to the input of a given cell. These two items together define an update rule for the state of each cell. If we the neighborhood includes all cells, and the transition function is an analog sigmoidal function, then we have a fully connected neural net. On the other hand, if the transition function is a simple mod(2) binary operation, and the neighborhood is a set of nearest neighbors, we have a classical cellular automata. These two extremes represent the range of possible massively parallel computation (on a chip). The intermediate combination of using the analog sigmoidal function and a limited neighborhood gives the cellular nonlinear network (CNN) of Chua [6]. Cellular automata have been used in image processing, and in simulation of complex transport. We have used a restricted neighborhood in layered neural networks, and a binary switching function, to design cellular chips [7]. These cellular systems can implement general purpose computation. *The task is to ascertain whether novel architectures and novel devices can be integrated to produce a continued increase in functionality in integrated circuits.*

The matrix **W**, describing how the neighborhood connects to the input of each cell is directly related to the state transition matrix of a Boolean decision tree. We have called these binary "neurons" Boolean McCulloch-Pitts neurons (BMPN) [7], and have implemented both analog and digital versions, with a 512 neuron chip being fabricated. The connection matrix **W** is a reduction of the actual state transition matrix **M**. This reduction proceeds through cut-set matrices, but allows us to talk about reconfiguration of the interconnections. Computation, in the sense of the Turing machine, proceeds through a set of state transition matrices, each of which arranges the possible Boolean states in the order necessary to perform a prescribed sub-computation. In principle, programmability can be achieved through the use of an activation function to set the weights that represent the elements of **W**.

It is entirely conceivable to create a general purpose, and massively parallel, binary computation machine on a chip, but in a different form than the normal von Neumann architecture that is currently the basis of microprocessors. More importantly, it means that it is possible to develop an optimum architecture for novel devices which may provide more operational capability than simple binary switching.

One novel device (for logic) which offers more than simple binary switching is the single-electron transistor, where a mod(2) function can be obtained. This has led several groups to consider multi-level logic with SETs, but this is not likely to be the best architecture. Instead, it is important to understand what architecture may be suitable for devices such as the SET, and how novel quantum wires and dots will fit into this architecture. Only then will continued scaling beyond the 22 nm node be realizable by jumping to a different technology trajectory in Fig. 1.

References

[1] M. J. Gilbert and D. K. Ferry, submitted for publication.

[2] D. K. Ferry and W. Porod, Superlatt. Microstruc. **2** (1986) 41.

[3] M. Yazdani, D. K. Ferry, and L. A. Akers, IEEE Circ. Dev. Mag. **13** (no. 2), (March 1997) 28.

[4] A. W. Burks, Ed., *Essays on Cellular Automata* (1970).

[5] E. F. Codd, *Cellular Automata* (1968).

[6] L. O. Chua and T. Roska, *Cellular Neural Networks and Visual Computing: Foundation and Applications* (2002).

[7] D. K. Ferry, R. O. Grondin, L. A. Akers, and L. C. Shiue, in *Frontiers of Computing Systems Research*, Vol. 1, Ed. by S. K. Tewksbury (1990) 47.

[8] S. M. Ramey and D. K. Ferry, Semicond. Sci. Technol. **19** (2004) S238.