Nanotechnology in the Development of Future Computing Systems

Toshishige Yamada and M. Meyyappan
NASA Ames Research Center
Center for Nanotechnology, MS 229-3
Moffett Field, CA 94035 USA
Ph: 650 604 2616 Fax: 650 604 5244
email: meyya@orbit.arc.nasa.gov
http://www.ipt.arc.nasa.gov

Abstract

Nanoelectronics based on carbon nanotubes and organic molecules as the conducting material has been receiving increasing attention as an alternative to silicon CMOS based computing technology. These new device technologies require development of novel circuits and architectures to be technically feasible and economically viable. This article provides an overview of this subject along with opportunities and challenges.

Introduction

Silicon CMOS technology is the backbone of the computer industry which has been the case since the invention of integrated circuits. The industry witnessed doubling of the chip density nearly every two years for the past decade. This trend, commonly known as the Moore's law, is expected to continue for a few more generations lasting another decade or so. At that point, while silicon CMOS based computer chips will continue to be made, the relentless scaling seen until now may not be possible. This scenario may be likely somewhere in the sub-50 nm feature scale regime. Even in the quest to reach that level, there are substantial technological barriers according to the International Technology Roadmap for Semiconductors (ITRS) [1]. These barriers include lithography, low-k dielectric materials, interconnects to name a few.

In recent years, the research community has focused on several alternatives including carbon nanotube based nanoelectronics, molecular electronics based on conducting organic molecules, single electron transistors, and quantum computing, to name a few of the most popular technologies. Here, a brief overview of carbon nanotube based devices is provided along with our views on the circuit and architecture issues for nanoelectronics.

Carbon Nanotube based Nanoelectronics

A carbon nanotube (CNT), which is an elongated fullerene, can be metallic or semiconducting depending on chirality and diameter [2]. The bandgap is given by

$E_g \quad 2 \, {}_{o}a_{c\;\;c}/d$ where $a_{c\;\;c}$ is the C-C bond length, d is nanotube diameter and γ_o is the nearest neighbor hopping parameter. The electronic properties are also affected by mechanical deformation and structural defects. The intriguing electronic properties of carbon nanotubes provide an opportunity to create metal-semiconductor and semiconductor-semiconductor junctions leading to functional devices.

In CNT based nanoelectronics, several groups have demonstrated operation of CNT-field effect transistors (FETs) [3-5]. Typically, a nanotube grown elsewhere was transplanted across a pair of gold electrodes [3, 4] serving as source and drain and in contact with a SiO_2 layer. A backgate at the bottom silicon substrate was used to modulate the current flow. Devices fabricated in this manner operated at room temperature and showed that the drain current (I_D) vs. drain voltage (V_D) behavior at a fixed gate voltage (V_G) did not show any saturation at large biases unlike in conventional silicon CMOS. In contrast, the drain current showed a saturation as a function of gate voltage at a given drain bias, again unlike in silicon CMOS. The threshold voltage was also large around several volts. Analyzing these CNT-FETs, Yamada [6] found that the metal electrode-nanotube contact influences the transport greatly and the saturationless I_D vs. V_D behavior indicates transport in the weak-localization regime in the absence of carrier-carrier scattering so that pinchoff cannot occur. To compensate for this behavior in digital applications, he suggested nanotube transistor designs to maximize the transconductance. Saturation in the I_D vs. V_G curves, Yamada concluded, reflects the one-dimensional nature of nanotube state density.

Following these early demonstrations, Liu et al. [5] integrated p-type and n-type nanotube transistors to create an inverter (see Fig. 1). They used chemical vapor deposition (CVD) to grow the nanotube bridging the source and drain and in contact with the SiO_2 layer. The same backgate configuration as in refs. 3 and 4 was used. Potassium doping was used to

Proceedings of the Application-Specific Systems, Architectures, and Processors (ASAP'03)
ISBN0-7695-1992-X/03 $17.00 © 2003 IEEE

IEEE COMPUTER SOCIETY

create n-type doping while the p-type device was undoped. Fixing the drain bias at 2.9 V, they swept the gate voltage from 0 to 2.5 V to demonstrate the transfer characteristics of the inverter.

Properties of nanoFETs vs. macroscopic FETs

In the present silicon-based technology, FETs play a central role in electronics. For this reason, current nanodevice research is focusing on the FET scheme. Here we compare the macroscopic FETs and nanoFETs in view of how the channel material would influence the device characteristics. Since the FETs are genuine three-terminal devices, we need to discuss the characteristics in response to the drain voltage and those in response to the gate voltage independently.

The drain voltage characteristics are strongly dependent on the channel material in macroscopic FETs, while they are mostly independent of the channel material in nanoFETs. In a macroscopic FET, the drain current as a function of drain voltage is proportional to the carrier mobility [7]. The mobility is proportional to the mean free time of a carrier traveling without being scattered, and the effective mass of the carrier. One of the main scatterers is phonon, and the phonon properties are different from material to material, reflecting the bulk material property difference. The effective mass represents the band structure of the bulk material and is again material-dependent. Adopting different channel materials will certainly result in different characteristics. However, in nanodevices with an ideal electrode contact, the drain current as a function of drain voltage is determined by a transmission coefficient in the Landauer-Butticker view [8]. In the limit of an ideal nanoFET where carriers can run ballistically from source to drain, the transmission coefficient is unity, and there is no difference among different channel materials. In reality, it is challenging to have an ideal electrode contact as seen generally at the metal-semiconducting nanotube interface [9]. The transmission coefficient is reduced at the contact in the example of a CNT-FET [3-5]. Depending on the channel material, the feasibility of taking a contact differs, and in this sense, the characteristics will depend on the channel material. It has to be noted, however, that the bulk channel material properties do not play any role in this discussion.

The gate voltage characteristics present an interesting situation for the macroscopic FETs and nanoFETs. In both cases, the problem is reduced to how effectively we can induce carriers in the channel using the gate capacitance. Thus, the characteristics are determined by the thickness and the dielectric constant of the insulating layer, which is an environmental structure outside the channel and this situation is similar in a gated nanotube Schottky diode [10]. Ironically, this has no direct connection to the channel material itself. Depending on the choice of the channel material, the feasibility of preparing the insulation layer may differ. There is no essential difference for the macroscopic and nanoFETs in this aspect.

In nanoFETs, the different channel material will result in different contact properties and lead to different drain voltage characteristics. The bulk material properties do not influence the drain characteristics in nanoFETs. This is in sharp contrast to the macroscopic FETs, which show a significant bulk channel material dependence. The gate voltage characteristics do not directly reflect the channel material properties in both the macroscopic and nanoFETs.

The forgotten benefit of nanoFETs - inherent freedom from the short channel effects

When Shockley developed the FET [11], he demanded that the FET channel should be very thin compared with the channel length, so that the variation of the electric field along the channel is much less than the corresponding variation perpendicular to the channel. In other words, the dominant electrostatic problem is in the perpendicular direction to the channel, and carriers are securely confined to the thin channel so that the distance to the gate electrode is minimal. Thus the on state and the off state of the transistor are very well distinguishable. A largest possible current ratio for on and off states is one of the basic requirements for FETs. In order to achieve this, we need to place carriers right below the gate electrode. Confined in a thin channel layer, carriers are only minimally away from the gate electrode and their concentration is most effectively controlled by the gate voltage. This would result in excellent transistor performance.

As the channel length shrinks, this situation does not hold any more, and the electric field variation near the source and the drain cannot be negligible. The carriers are no longer confined to a thin layer below the gate electrode and are distributed deeper into the substrate near the source and the drain. A significant portion of the carriers is away from the gate electrode. This will cause a so-called short channel effect. In this situation, the gate voltage cannot effectively control the carrier concentration and the transistor does not shut off well. Even if we apply a gate voltage to shut off the transistor, there is still significant drain current, and this degrades the transistor performance. This is often referred to as the reduction of threshold voltage and is one of the serious problems the current CMOS technology faces. We are forced then to engineer the channel doping to suppress these unwanted short channel effects, but such doping control is highly challenging.

The fundamental solution is to adopt a transistor channel material that is, from the beginning, two-dimensional or one-dimensional so that the channel thickness is minimal. Nanodevices based on molecules, CNTs, or DNAs are fortunately quasi-one-dimensional, and will be potentially quite advantageous in inherently suppressing the short channel effects. In fact, CNT-FETs have shown excellent performance which is already comparable to that of the state-of-the-art CMOS technology, and this is largely due to this "thin layer effect" through the adoption of a thin nanotube for the channel.

IEEE COMPUTER SOCIETY

This article is the first to emphasize this important, but forgotten benefit of nanotechnology in electronic device applications. Nanodevices are inherently free from the short-channel-effects, and this serious problem will not be faced in the future. Despite many other problems that may arise in the future, it is quite meaningful to pursue nanotechnology in this context, in addition to the apparent benefit of the ultimate small size.

The two-terminal vs. three-terminal: inherently two-terminal nanodevices need a new circuit scheme

In electronics applications, we pay a special attention to input and output electrodes. When a device has electrically independent input and output electrodes, it is called three-terminal having an input electrode, an output electrode, and a ground electrode. The electric isolation of the input and the output electrodes is essential, because this guarantees that the input and the output signals are not mixed. Because of this isolation, in series connection of these devices, the signal can transmit only in the designed direction. Consider a series connection of device A, device B, and device C such that the output of A drives the input of B and the output of B drives the input of C. In this case, if we give a signal to the input electrode of device B, the signal transmits only to C, and it will not bounce back to A because the output and input electrodes are isolated in A. Therefore, the signal flow from A to B to C is established. A two-terminal device has only two electrodes, where one is for ground and the other is for input/output. In the series connection of device A, device B, and device C, if we give a signal to the input electrode of device B, the signal will transmit to A as well as C because the output and the input electrodes of A is not isolated. The signal flow is not well established, and this is quite disadvantageous in circuit applications. As is obvious here, even if a device has three electrodes, if the input and the output electrodes are not electrically isolated, it is not three-terminal.

An example of three-terminal device is a MOSFET. It has three electrodes, and in the most basic inverter circuit application, the gate electrode receives an input signal, the drain electrode gives an output signal, and the source electrode is connected to the ground. The gate electrode and the drain electrode are electrically isolated via a silicon dioxide layer, which is an excellent insulator, and there is no possibility for the input-output signal mixture. A Schottky diode is two-terminal, because it has only two-terminals, one is to be used for input/output and the other is to be used for ground. A Josephson junction diode is another example of two-terminal devices. In these diode examples, special circuitry has to be built to prevent a reverse, unwanted signal flow (from device B to device A in the example above).

Extensive circuit libraries suitable for three-terminal devices are readily available and in fact the current CMOS technology relies on them. Thus, we can immediately build circuitry with any new devices as long as they are three-terminal. We can basically drop in the new devices for the present MOSFETs. The situation is not this straightforward for two-terminal devices, because we need to develop different circuit libraries than what we are using now. However, we do not have to start from scratch. Historically, at least a few times, circuit libraries for two-terminal devices were developed. In the 1950s when semiconductor three-terminal devices were not very reliable but semiconductor diode devices were, "diode logic" was proposed which provided a circuit scheme based on the two-terminal devices only [12]. In the 1970s, Josephson junction diodes were studied extensively. In these studies, it was shown that "quasi-three-terminal gates" having independent input and output electrodes would be created by connecting diode elements [13]. Two terminal devices were studied whenever new diodes based on new technology were invented.

In nanodevices, we often use molecules, CNTs, or DNAs for the device channels. All these structures are one-dimensional having two ends. Thus, without any special effort to place a third electrode, the device is inherently two-terminal. The proposal [14] by the HP group to place a foundation in the diode logic scheme is based on this understanding.

Alternatives to CMOS Architecture?

The early results for nanoelectronics devices using novel conducting channel materials are interesting but there is a long road ahead prior to realizing viable computer technologies based on carbon nanotubes or organic molecules. There are several reasons for this. Introduction of revolutionary technologies, particularly a disruptive technology, is always a long arduous process. In the present case, no novel processing schemes appear to be in sight. While the conducting channel material is new, the processing scheme is almost the same as in current silicon CMOS practice. It is well known that one of the serious challenges to CMOS scaling now is the cost of new fabs. Novel material based nanoelectronics at the end of the silicon roadmap, if it were to use somewhat identical processing scheme, would perhaps price itself out of existence. In that regard, the most promised outcome/benefit of nanotechnology – novel bottomup processing schemes, self assembly, etc. – is critical to develop a viable nanoelectronics based on nanotubes or molecules.

In addition to the need for cost effective processing, novel architectures are also highly desirable. Indeed, if a feasible new architecture is developed, that may drive the innovation of processing techniques in the right direction. Earlier we pointed out the differences between nano- and macro- FETs, some advantages of nanoFETs in terms of short channel effects,

IEEE COMPUTER SOCIETY

and robust input-output isolation in three terminal devices like CMOS. All of these do not necessarily mean that we should continue to rely on CMOS architecture; in addition to expensive processing schemes there are power considerations as well.

Figure 2 shows isoenergy plots for CMOS devices, in their delay time vs. power dissipation per gate [15]. As expected, as the delay time decreases, the power dissipation increases. For CMOS devices, the isoenergy plot improves with each generation and it currently stands at 10^{-16} J. Interestingly, this coincides with the estimated isoenergy line for the brain cell. For a CMOS architecture, 16^{-16} J/operation would result in 10^{-7} J/s or 10^{-7} W for a 1 GHz clock operation. The number of gates is expected to be 10^{8} between now and 2005 which gives a power dissipation of tens of Watts. This increasing level of power dissipation is a serious concern. Biological cells, though slow compared to CMOS, compensate for the difference through architecture of the system. It is desirable then to develop alternatives to the current CMOS architectures, in the context of the emerging nanoelectronics, to push the operating regime towards the lower left corner of Fig. 2.

This means a low temperature operation of nanodevices, and in an advanced stage of nanoelectronics when the heat dissipation is becoming a problem, a low temperature operation will become mandatory. In digital electronics applications, we need to store data representing "0" and "1". In any kind of device with storage functionality, we are required to use an energy to convert "0" to "1", or "1" to "0", otherwise the stored data may be lost due to a thermal perturbation without our notice. In order to prevent such malfunction, the data conversion energy per bit, called barrier energy, must be much larger than the thermal energy. The current CMOS architecture uses a barrier energy of 10^{-16} J, which is still orders of magnitude larger than the thermal energy of 10^{-21} J at a room temperature. We can reduce this barrier energy in an early stage of nanotechnology by achieving a radical device miniaturization. In an advanced stage of nanoelectronics when the thermal limit is met, we will have to abandon the room temperature operation and pursue a low temperature operation.

Early proposals of novel architecture include a CNT neural tree [16] which is a 3D network of chemically interconnected nanotubes with nodes at the junction serving as devices and switches. Figure 3 shows a CNT dendritic neural tree which has four levels of branching structure and is made up of 14 Y-junctions. The Y-junction is a novel three branch formation [17] which has been observed in CNT growth [18,19]. The branching and controlling of signals at these junctions could be made to mimic biological neural networks. It is envisioned [16] that this neural tree can be trained to perform complex signal processing in computing applications.

References

[1] http://www.itrs.net/ntrs

[2] R. Saito, M.S. Dresselhaus, and G.S. Dresselhaus, Physical Properties of Carbon Nanotubes (World Scientific, River Edge, 1998).

[3] S.J. Tans, A.R.M. Verschueren, and C. Dekker, “Room temperature transistor based on a carbon nanotube,” *Nature*, Vol. 393, pp. 49-52, 1998.

[4] R. Martel, T. Schmidt, H.R. Shea, T. Hertel, and Ph. Avouris, “Single and Multiwall Carbon Nanotube Field-Affect Transistors,” *Appl. Phys. Lett.*, Vol. 73, pp. 2447-2449, 1998.

[5] X. Liu, C. Lee, C. Zhou, and J. Han, “Carbon Nanotube Field Effect Invertors” *Appl. Phys. Lett.*, Vol. 79, 3329-3331, 2001.

[6] T. Yamada, “Analysis of Submicron Carbon Nanotube Field-Effect Transistors,” *Appl. Phys. Lett.*, Vol. 76, pp. 628-630, 2000.

[7] S.M. Sze, Physics of Semiconductor Devices, 2nd ed. (Wiley, New York, 1981); Y. Taur and T.H. Ning, Fundamentals of Modern VLSI Devices (Cambridge, New York, 1998).

[8] Y. Imry, Introduction to Mesoscopic Physics (Oxford, New York, 1997); M. Lundstrom, Fundamentals of Carrier Transport, 2nd ed. (Cambridge, New York, 2000).

[9] T. Yamada, “Modeling of Electronic Transport in Scanning Tunneling Microscope Tip Carbon Nanotube Systems,“ *Appl. Phys. Lett.*, Vol. 78, pp. 1739-1741, 2001.

[10] T. Yamada, “Modeling of Kink-Shaped Carbon Nanotube Schottky Diode with Gate Bias Modulation,” *Appl. Phys. Lett.*, Vol. 80, pp. 4027-4029, 2002.

ISBN0-7695-1992-X/03 $17.00 © 2003 IEEE

IEEE COMPUTER SOCIETY

[11] W. Shockley, "A Unipolar "Field-Effect" Transistor", *Proceedings of the I. R. E.,* pp.1365-1376, 1952.

[12] P.W. Becker, “Static Design of Transistor of Diode Logic,” *IRE Trans. Circuit Theory*, Vol. 8, pp. 461-467, 1961.

[13] T. Gheewala, “A 30-ps Josephson Current Injection Logic (CIL).” *IEEE J. Solid-State Circuits*, Vol. SC-14, pp. 787-793, 1979.

[14] J. Birnbaum and R. Stan Williams, “Moore’s Law and the Future of Computing,” *Phys. Today*, Vol. 53 (1), pp. 38-42, 2000.

[15] E. Takeda, *Microelectronics Rel.*, 1997.

[16] D. Srivastava, M. Menon, and K. Cho, “Computational Nanotechnology with Carbon Nanotubes and Fullerenes,” *Comp. Eng. Sci.*, pp. 42-54, 2001.

[17] M. Menon and D. Srivastava, “Carbon Nanotube based Molecular Electronic Devices,” *J. Mater. Res.*, Vol. 13, pp. 2357-2362, 1998.

[18] C. Papadopoulos, A. Rakitin, J. Li, A.S. Vedeneev, and J.M. Xu, “Electronic Transport in Y-junction Carbon Nanotubes,” *Phys. Rev. Lett.*, Vol. 85, pp. 3476-3480, 2000.

[19] C. Satishkumar, P.J. Thomas, A. Govindraj and C.N.R. Rao, “Y-Junction Carbon Nanotubes,” *Appl. Phys. Lett.*, Vol. 77, pp. 2530-2532, 2000.

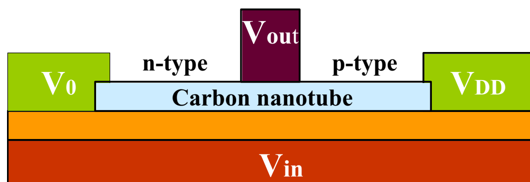

Figure 1. Carbon nanotube field effect inverter

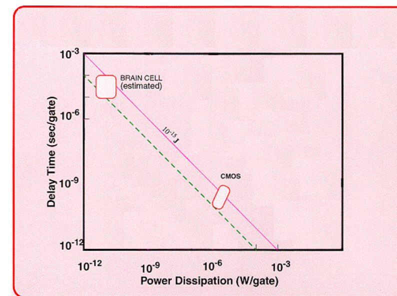

Figure 2. Switching time vs. power dissipation

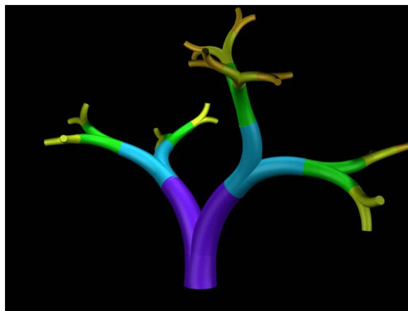

Figure 3. Carbon nanotube neural tree

Proceedings of the Application-Specific Systems, Architectures, and Processors (ASAP’03)
ISBN0-7695-1992-X/03 $17.00 © 2003 IEEE

IEEE COMPUTER SOCIETY