The Future of CMOS: More Moore or the Next Big Thing?

Wieslaw Kuzmicz
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
Warsaw, Poland
wbk@imio.pw.edu.pl

***Abstract*—This paper discusses the industrial and research status of CMOS and CMOS-like commercial and emerging technologies. Effects of scaling on transistor cost, max. system complexity and performance are discussed. Performance of state-of-the-art CMOS VLSI systems is power-constrained. To discuss various existing and emerging technologies, a model of an abstract, technology-independent ideal switch is proposed. It is shown that in the power-constrained scenario the maximum performance (in terms of clock speed) is limited by the static I_{off} current, and the I_{on} current (i.e. the current flowing when the device is switched on) is of secondary importance. CMOS and CMOS-like technologies: bulk CMOS, FDSOI, FinFET, tunnel FETs, junctionless FETs, nanowire FETs, and non-Si devices (e.g. carbon-based devices) are briefly presented and their technology readiness level is estimated. The main conclusion is that Si-based CMOS will remain the key VLSI technology for the next decade or longer, and older technology nodes (65 nm – 350 nm) will not be abandoned.**

***Keywords*—component; CMOS; FDSOI; FinFET; tunnel FET; junctionless FET; nanowire FET; graphene FET; GaN; InAs**

I. INTRODUCTION

The trend observed by Gordon Moore in 1965 [1], later named "Moore's Law", became a self-fulfilling prophecy and served as a guide for long-term planning of industrial R&D [2]. Now the transistor gate length has been reduced to 14 nm and below, the number of transistors in state-of-the-art high performance microprocessors exceeds 10^9. However, the thermal constraints limit the clock speed to 3…4 GHz and lithography challenges make further downscaling of device dimensions (a.k.a "More Moore") extremely difficult and expensive. What's next? The bulk CMOS technology has reached its limits around 28 nm process node, two new device architectures (FinFET and FDSOI) have been successfully commercialized and a number of alternative technologies, devices and semiconductor materials are being investigated. In this overview the limits of "More Moore" are discussed from the viewpoints of device cost, raw device performance, device density (i.e. number of transistors per unit area) and power-limited system performance. We introduce the concept of an abstract, technology-independent switch, which allows to show in a simplified but general way which key device properties affect maximum clock frequency and how.

Discussions in this paper are limited to CMOS and CMOS-like logic. The SRAM, DRAM, nonvolatile memory technologies etc. are beyond the scope of this paper.

This work has been supported by the National Centre for Research and Development in the frame of PBS program under Research Project "VESTIC", project no. 177244.

II. MOORE'S LAW IN THE PAST AND NOW

A. Moore's Law in the Past

The original observation published by Gordon Moore in 1965 was formulated in terms of device density and cost: "The complexity for minimum component costs has increased at a rate of roughly a factor of two per year" [1]. The measure of "complexity" was defined as the number of devices in an integrated circuit. The trend observed by Moore, named "Moore's Law", was continued when CMOS technology replaced other IC technologies. Every two to three years a new generation of CMOS technology appeared featuring smaller transistors. Minimum physical gate length became widely accepted as the label of the technology generation also called "technology node".

There were three main driving forces behind Moore's Law:

- cost reduction: smaller transistors were cheaper,
- performance increase: transistors with shorter gate lengths were inherently faster as digital switches,
- higher system complexity: smaller transistors -> more transistors per chip.

B. Are smaller transistors cheaper?

They used to be. For each new technology node cost of unit area of processed silicon increased, but number of devices per unit area increased faster, hence the cost per transistor decreased. However, below 20 nm the manufacturing costs increase at a much faster rate. One of the main reasons is lithography. The most advanced litho scanners still use 193 nm ultraviolet light. Below 20 nm many layers require multipatterning – two or more masks and litho steps per layer. This dramatically decreases fab throughput and increases costs. As a result, "the lowest logic gate cost is for 28 nm node" [3]. Multipatterning will not be needed when extreme UV lithography scanners (13.5 nm UV light) achieve production-

ready status. At the moment of this writing EUV scanners are available but their throughput is too low for a commercial fab. So far they can be used for process development only. The problem is in insufficient intensity of the UV light. At 13.5 nm reflective optics must be used (the masks are also reflective). The reflectivity of the mirrors is about 70% at most. After traveling between 10 or more mirrors (and reflective mask) the light intensity is reduced to several percent of the intensity at the source. As a result, wafer exposure time is very long. EUV light is produced by plasma produced by pulses of powerful CO_2 laser light colliding with tiny droplets of molten tin. So far it proved to be very difficult to boost the EUV light intensity at the source to the level of 250 W or more - sufficient for production-ready scanner - but in principle it not impossible. However, production EUV scanners will be much more expensive than existing 193 nm scanners. Will the transistor cost go down? It is difficult to predict.

C. *Are smaller transistors faster?*

They used to be. According to the classical „constant field" scaling theory (Dennard *et al.* [4]) device dimensions are scaled down together with the drain voltage. In this case, if the dimensions and drain voltage are divided by a factor k>1 and the channel doping is multiplied by k^2, the gate delay time also is reduced by k, and the dynamic (switching) power density does not change. Another scaling scenario is the „constant voltage" scaling, when all dimensions are divided by k>1, the channel doping is increased by k, but the drain voltage remains constant. In this case gate delay is reduced by k^2, but the dynamic power density increases dramatically – as k^3. In the early days of CMOS technology (down to 500 nm gate length) the maximum drain voltage was kept constant (5 V). From 500 nm down to 120 nm the drain voltage had to be reduced proportionally to the gate length: 5V->3.3V->2.5V->1.8V->1.2V - constant field scaling. Increasing channel doping resulted in somewhat decreasing channel mobility but still gate delay decreased with decreasing gate length. Below 120 nm constant field scaling is no longer possible, the drain voltage is only slightly reduced with every new technology node. The consequences are dramatic. Mobility decreases not only due to increasing channel doping but also because of strong electric field in the channel. To make the transistors faster, mobility enhancement techniques such as stress in the channel and Ge doping are used (note that this makes the process more complex and expensive). However, even if the gate delay decreases, the system clock frequency cannot increase because the power density limit (power per chip unit area) would be exceeded.

D. *Power-limited performance*

In devices with extremely short gates static power adds to the dynamic (switching) power. The main component of the static power consumption is the subthreshold current, i.e. the current I_{off} that flows when the gate voltage is zero. This current is proportional to $\exp\left(-qV_T/nkT\right)$, where V_T is the threshold voltage and n is a constant coefficient somewhat higher than 1. Its value determines the subthreshold slope, i.e. the slope of the I_D-V_{GS} characteristics in semilogarithmic coordinates. The "ideal" slope for n=1 is 60 mV per current decade at room temperature. While in the old technologies the typical threshold voltage was of the order of 1V, in transistors with extremely short gates and maximum drain voltages slightly below 1V the threshold voltage must be reduced to 0.3 … 0.4V. Moreover, in MOS transistors of traditional „bulk" architecture the n coefficient increases with decreasing device dimensions. V_T reduction and increase of n leads to faster than exponential increase of I_{off} when the gate length is reduced – see Fig.1.

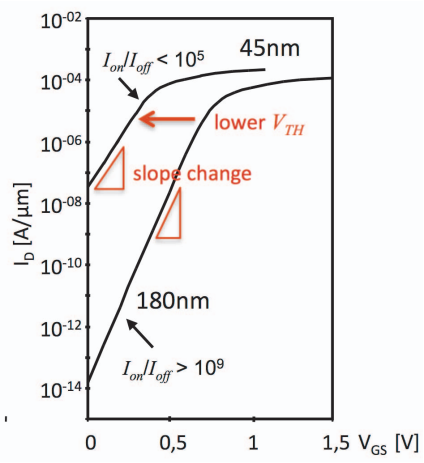

Fig. 1. Example of I_D vs. V_{GS} characteristics for 180 nm gate length and 45 nm gate length. Lower V_T shifts the characteristics to the left while increasing n decreases the slope. Both increase the I_{off} current.

At gate lengths below 65 nm the static current consumption may be comparable with its dynamic counterpart, „stealing" the power needed for switching: when the maximum power density is reached, more static power means less dynamic power, i.e. lower clock frequency. There are circuit and system-level countermeasures reducing system static power, such as power and clock gating, multiple cell libraries (low power, high performance) etc., but they make system design much more complex and do not solve the static current problem completely.

E. *FinFET and FD-SOI*

Two new MOS device architectures – FinFET and FD-SOI – achieve lower I_{off} current by means of better electrostatic control of the channel, which is formed in a thin Si "fin" or a thin Si layer on a SOI substrate (Fig. 2).

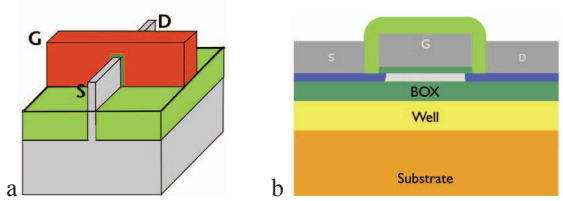

Fig. 2. FinFET (a) and FD-SOI (b) transistors

Better electrostatic control of the channel leads to steeper subthreshold slope, i.e. *n* closer to unity. As a result, either lower power consumption or higher performance can be achieved. FinFET technology is more appropriate for large digital systems while FD-SOI is more universal. Moreover, FD-SOI transistors do not need any doping in the channel. This significantly reduces number of masks in comparison with traditional "bulk" transistors. Although SOI wafers are more expensive than ordinary Si wafers, FD-SOI is not more expensive than "bulk" or FinFET technology.

Without FinFET and FD-SOI shrinking transistor gate length below 28 nm would make little sense – no cost reduction and performance limited by maximum power density, not by raw device speed. FinFET and FD-SOI helped to extend the Moore's Law. Major IDMs and foundries offer technologies with gate length below 20 nm and promise to go down to 7 nm and later 5 nm. Commercial viability of technologies below 10 nm is, however, questionable. ITRS [2] predicts that transistor shrinking will stop at 10 nm (Fig. 3).

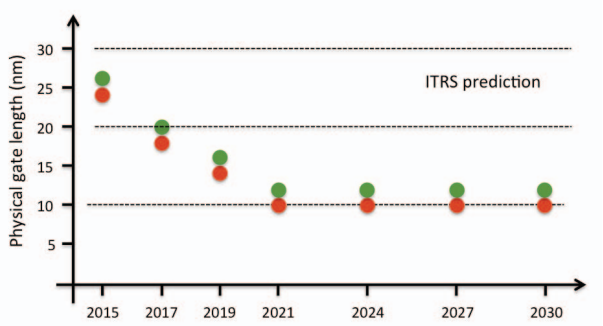

Fig. 3. Minimum physical gate length in commercial CMOS digital technologies – ITRS prediction. Red: "high performance" devices, green: "low power" devices

F. Area scaling

While transistor gate length scaling below 10 nm may not make sense, other ground rules such as metals half-pitch, contacted poly half-pitch, fin spacing, spacer width etc. (Fig. 4) may be scaled down, and as a result the transistor density will increase. The goal is to have more devices per chip, in order to design SoCs with more advanced architectures and functionality.

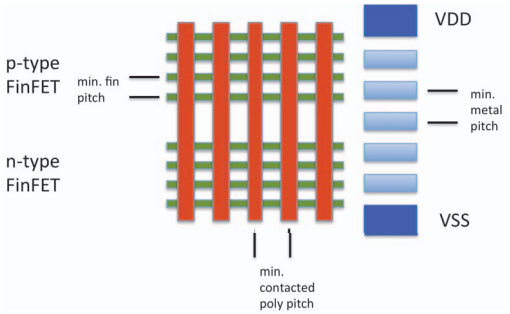

Fig. 4. Some layout rules that can be scaled down

ITRS predicts (Fig. 5) that some of the rules may be scaled down to approx. 5nm.

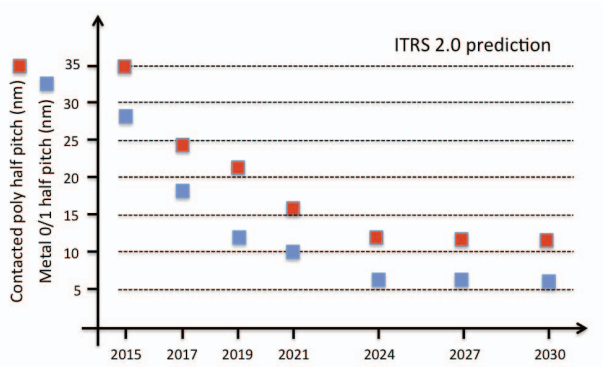

Fig. 5. Scaling of some of the layout rules – ITRS prediction

Recently Intel proposed [5] to introduce new "node" metric based on average transistor density:

$$\text{No. of transistors/mm}^2 = 0.6\,\frac{\text{NAND2 tr. count}}{\text{NAND2 cell area}} + 0.4\,\frac{\text{Scan FF tr. count}}{\text{Scan FF cell area}}$$

Note that with this definition we return to the original idea of Moore's Law, which was formulated in terms of device density, not gate length.

G. CMOS vertical extensions: 2.5D, 3D

The idea of vertical integration is not new. First attempts are older than 10 years. Two examples are: SOI-based complete radiation imaging chip with pixel matrix in the bulk and CMOS readout electronics over the BOX layer (IET Warsaw) and autonomous sensor "cubes" – 3D stacks of 2D functional modules, most of them - CMOS chips.

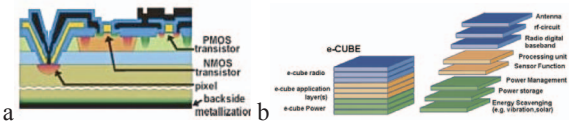

Fig. 6. Early 3D structures: radiaton imaging chip (a) and autonomous sensor "ecubes" (b). Picture sources: (a) IET Warsaw, (b) eCUBES project Web page

Advantages of vertical integration are: very short interconnections, low interconnect capacitance, higher performance, lower power consumption and small footprint of the complete "System in Package" (SiP). Fig. 7 shows examples of SiP structures in which CMOS chips are located on a passive interposer (also called redistribution layer), which in turn can be connected to a PCB via a BGA-type interconnections. This arrangement is often called "2.5D integration". Fig.8 shows an example of true 3D integration: stacked DRAM chips with Through Silicon Via (TSV) interconnections.

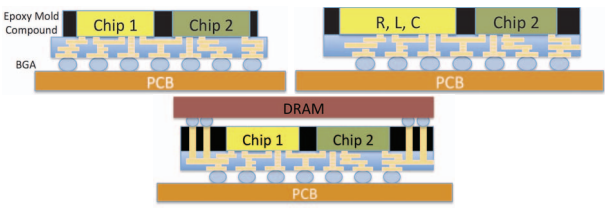

Fig. 7. System in Package: various examples of 2.5D integration

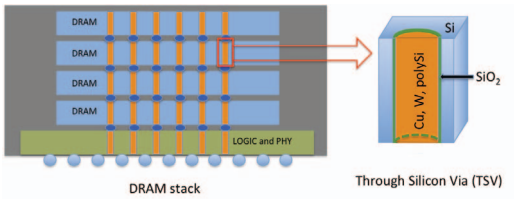

Fig. 8. 3D integration: stacked DRAM chips with TSV vertical interconnections

H. Current industrial landscape

At the moment of this writing the most advanced CMOS technologies that are available on the market have gate lengths between 10 nm and 20 nm. In spite of ITRS prediction major foundries and IDMs claim that they will have 7 nm and 5 nm fully qualified, high yield processes, FinFET and/or FDSOI, within the next few years. However, these technologies will coexist with older ones. 28 nm technologies of various flavors will probably dominate in the next 5 to 7 years. Moreover, 17 new fabs for 200 mm wafers and 120 nm – 180 nm processes are now being built or planned because there is still high demand for older CMOS technologies, for applications that require neither extremely high performance nor ultra-low power.

SiP 2.5D integration is used mainly for mobile devices where space is very limited. 3D integration has been demonstrated for DRAM modules.

III. A Glimpse to the Future

Transistor gate length is or will soon be finished, and area scaling will be finished within the next 5 to 7 years. The question is: how to continue progress in micro/nanoelectronic systems? What options exist? New device architectures? New kinds of semiconductor devices? New semiconductor materials?

A. What are we looking for?

Let us first find the most important properties of new devices, which might potentially replace Si-based MOSFETs. In all solid-state logic circuits that have found practical applications so far the computational variables (i.e. physical quantities that represent logical values) are: voltage, current and charge (VCC). The considerations in this paper are limited to this class of devices. For a broader overview of possible post-CMOS devices the reader is referred to [6].

For VCC-based devices a simple, technology independent model that will be called abstract switch can be defined. The abstract switch is a device located on a substrate that dissipates power P generated by the switch and for which the maximum power density is $P_{d\max}$. The power P is dissipated across the area A, which will be called device thermal footprint (Fig. 6). The switch is in the ON state during the time τ and in the OFF state during the idle time $t - \tau$. In the ON state the constant current I_{on} charges the capacitance C from 0 to the supply voltage V_{DD}: $I_{on}\tau = Q = CV_{DD}$. In the OFF state the current equals I_{off}.

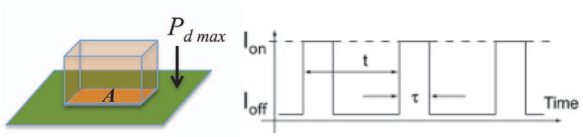

Fig. 9. The abstract switch and its operation

The energy E consumed during the clock cycle t is:

$$E = I_{on}V_{DD}\tau + I_{off}V_{DD}t = CV_{DD}^2 + I_{off}V_{DD}t \tag{1}$$

The average power $P = E/t$ dissipated across the area A must not exceed $P_{d\max}$

$$\frac{P}{A} = \frac{CV_{DD}^2/t + I_{off}V_{DD}}{A} \le P_{d\max} \tag{2}$$

This condition determines the minimum clock cycle time $t_{\min}$ and the maximum clock frequency $f_{\max}=1/t_{\min}$:

$$t_{\min} = \frac{CV_{DD}^2}{P_{d\max}A - I_{off}V_{DD}} \tag{3}$$

$$f_{\max} = \frac{P_{d\max}A - I_{off}V_{DD}}{CV_{DD}^2} \tag{4}$$

The capacitance charge time τ is not longer than the cycle time t, hence the maximum clock frequency f_{max} can be obtained if the I_{on} current is high enough to deliver the charge Q to the capacitance C during the cycle time $t_{\min}$

$$I_{on} \ge I_{on\min} = \frac{CV_{DD}}{t_{\min}} = \frac{P_{d\max}A - I_{off}V_{DD}}{V_{DD}} \tag{5}$$

The maximum clock frequency f_{max}, although not directly related to the actual clock frequency of a digital system, is a good measure of raw *power-limited performance* of a switch. It depends on supply voltage V_{DD}, the load capacitance C (where part of it may be external to the device), the device thermal footprint A and the I_{off} current, but does not depend on the I_{on} current provided that I_{on} is not lower than $I_{on\min}$. This indicates that, when performance is power-limited, to maximize the performance the I_{off} current and supply voltage V_{DD} should be minimized. The maximum power density $P_{d\max}$ and the device

thermal footprint A should be maximized, i.e. the heat generated in the device should be dissipated as effectively as possible.

The I_{on}/I_{off} ratio is often treated as the most important figure of merit of semiconductor devices to be used in digital CMOS or CMOS-like circuits, while in fact the I_{off} value is more important when performance is power-limited.

B. *New device architecture: nanowire MOSFET*

A stack of nanowire FETs is the next step in evolution of FinFET technology: "Gate All Around" (GAA) (Fig. 10).

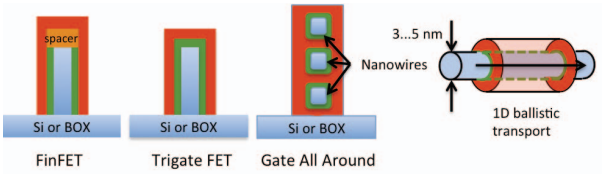

Fig. 10. FinFETs and nanowire transistors

Nanowire transistors (with wire dimensions of several nm) feature good electrostatic control of the channel (low I_{off}) and 1D ballistic transport (high I_{on}). Unfortunately heat dissipation can be problematic, especially for the upper wires in the stack. Fabrication process of the GAA stacks has been demonstrated [7] and it is expected that this device architecture will replace FinFETs in the next decade [2].

C. *New semiconductor device: tunnel FET*

The structure of a tunnel FET (TFET) is similar to a traditional planar MOS device, the difference is that source and drain are of opposite types (Fig. 11):

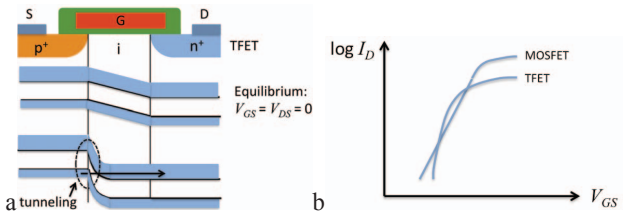

Fig. 11. Tunnel FET (a) and its drain current vs. gate voltage characteristics (b) compared with characteristics of a comparable MOSFET

The tunnel FET is switched on by applying gate bias so that electron accumulation occurs in the intrinsic region. At sufficient gate bias, when the conduction band of the intrinsic region aligns with the valence band of the p^+ region, electrons from the valence band tunnel into the conduction band of the intrinsic region and to the n+ region. Tunnel FET is a "steep slope" device, I_{off} current is very small, (see the drain current vs. gate voltage characteristics - Fig. 11b), much smaller than in comparable MOSFET. Unfortunately, I_{on} current is also small, too small for digital logic gates. Larger I_{on} current may be obtained in non-Si TFETs, in semiconductors with higher tunneling probability than in silicon.

Another problem with TFETs is that it is difficult to make a complementary pair needed for CMOS-type gates.

D. *New semiconductor devices: juntionless FET, VeSFET*

Junctionless FET is a gate voltage-controlled resistor. Source, drain and channel are of the same type (Fig. 12).

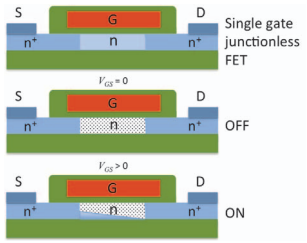

Fig. 12. Single gate planar juctionless FET

In the OFF state the channel is fully depleted due to the work function difference between the channel and the gate. Sufficiently high bias applied to the gate narrows the depleted part of the channel and the transistor is switched ON. In this state the width of the undepleted part of the channel and the drain current are modulated by the gate voltage. The characteristics of junctionless transistors are very similar to the characteristics of the ordinary MOSFETs. An important advantage of junctionless transistors is that they may have very low I_{off} currents.

Junctionless FETs of various architectures have been reported (see e.g. [8] for one example). One specific architecture of a junctionless FET is the twin gate vertical slit FET (VeSFET) [9]. Fig. 13 shows its unique architecture: the channel is in a narrow slit located vertically between two gates and four metal rods called pillars. Two of them serve as contacts to two gates, the other two are contacts to source and drain. The circuit is a regular array of VeSFETs (Fig. 14), where n-channel devices and p-channel devices can coexist, making CMOS-like gates possible. The mask shapes are lithography-friendly: circles and rectangles with rounded corners. Two symmetrical independent gates open the way to new circuit concepts, e.g. voltage applied to one gate may control the threshold voltage seen by the other gate. The array of VeSFET devices is inherently suitable for 3D integration without need for TSV connections. It is also worth noting that in the basic VeSFET cell other kinds of semiconductor devices can also be made, e.g. if there is no gate oxide layer, the device can be used as either a junction FET or a bipolar transistor.

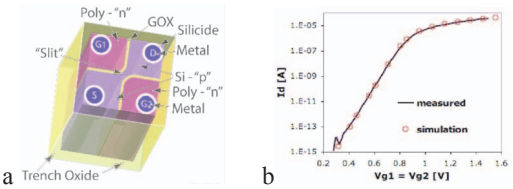

Fig. 13. VeSFET: twin gate vertical junctionless transistor (a) and its characteristics (b)

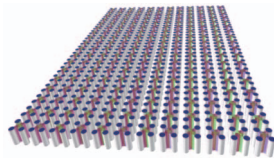

Fig. 14. VeSFET-based circuit: array of VeSFET cells, all have the same shape and dimensions, but each of them can be either n-channel or p-channel. Routing over the cells not shown.

Recently published simulation based study [10] suggests that VeSFETs might be the devices of choice for 7 nm node and beyond. Moreover, another simulation study [11] shows that VeSFET-based circuits have better vertical thermal conductivity than their CMOS counterparts.

E. Other materials and devices

Semiconductors other than Si include Ge, GaAs, GaN, InAs and graphene. All of them were tried as possible replacements for silicon. However, their properties make them less suitable for digital CMOS-like logic. A Ge-based nanowire transistors and CMOS inverters have been reported recently [12]. Germanium has higher carrier mobility than silicon but unfortunately narrower bandgap (0.66 eV). As a result, the maximum operating temperature for Ge-based devices is much lower than their Si-based counterparts.

GaAs, GaN, InAs have very high electron mobility but much lower hole mobility (e.g. for InAs: μ_n=40000 $cm^2/(Vs)$, μ_p=500 $cm^2/(Vs)$). Although they do have important applications (e.g. GaAs – RF devices, GaN - power devices), they are not suitable for complementary pairs of n-channel and p-channel devices.

Graphene has also been tried as the material for MOSFETs. However, graphene as such has zero bandgap. A MOS device with graphene-based channel could not be switched off. Graphene may, however, somewhat improve Si-based CMOS ICs, e.g. it has been suggested that a layer of graphene on top of Cu interconnections may reduce resistance and prevent electromigration. Bilayers of graphene and similar materials can, however, be used to make working transistors. This is confirmed not only by simulation [13], but also by experiment [14] – a simple digital circuit has been demonstrated with transistors based on MoS2 bilayer. This circuit, however, uses n-channel transistors only and digital gates are built with pulldown n-channel transistors and n-channel depletion-mode load devices. This is an old technique long ago abandoned because of large currents consumed by the gates.

Carbon nanotubes are another material for MOSFETs. Carbon nanotube based transistors and simple circuits have been demonstrated recently [15]. The contacts to source and drain are made of graphene layer. The authors of this work claim that both n-channel and p-channel devices can be made and that they can be up to 3 times faster than Si-based MOSFETs. However, the heat produced in the nanotubes will be probably difficult to dissipate effectively.

IV. Conclusions

Despite many attempts to find "The Next Big Thing" – a technology that would replace Si-based CMOS and offer either higher performance or lower power consumption or both, and at the same time would be manufacturable in high volumes at reasonable costs, it seems that Si-based CMOS will remain the key technology for logic gates for the next decade or two. Other device concepts and/or new materials, even if promising, are at this moment only at the "proof of concept" stage. Historical data indicates that time needed to go from this stage to the mature manufacturable product is of the order of 15 to 20 years. In this overview we do not consider data processing paradigms other than traditional "voltage, current and charge"-based digital circuits, such as spintronic devices, quantum computing, bio-inspired computing etc. Will any of them become someday "The Next Big Thing"? It is an open question.

References

[1] G. E. Moore, "Cramming more components onto integrated circuits", Electronics, Volume 38, Number 8, April 19, 1965

[2] International Technology Roadmap for Semiconductors ITRS 2.0, available at: http://www.itrs2.net, last accessed May 24, 2017

[3] H. Jones, "Semiconductor Industry from 2015 to 2025", International Business Strategies (IBS) whitepaper, 2015, available at: http://www.semi.org/en/node/57416, last accessed: May 23, 2017

[4] R. H. Dennard *et al*., "Design of ion-implanted MOSFET's with very small physical dimensions", IEEE J. of Solid-State Circuits, vol. SC-9, No. 5, Oct. 1974

[5] Presented at Intel's Manufacturing Day, 28.03.2017

[6] D. E. Nikonov and I. A. Young, "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking", Proc. of the IEEE, vol. 101, No. 12, December 2013, DOI: 10.1109/JPROC.2013.2252317

[7] C. Dupre et al., "15nm-diameter 3D stacked nanowires with independent gates operation: ΦFET**,** IEDM Tech. Dig., p.749, 2008, DOI: 10.1109/IEDM.2008.4796805

[8] J. P. Colinge *et al*., "Nanowire transistors without junctions", NatureNanotechnology, vol. 5, pp. 225-229, Feb. 2010, DOI: 10.1038/nnano.2010.15

[9] W. Maly *et al*., Twin gate, vertical slit FET (VeSFET) for highly periodic layout and 3D integration, Proc. MIXDES 2011, pp. 145-150, 2011

[10] P-L. Yang, T. B. Hook, P. J. Oldiges and B. B. Doris, "Vertical slit FET at 7-nm node and beyond", IEEE Trans. Electron Devices, vol. 63, No. 8, August 2016, DOI: 10.1109/TED.2016.2577629

[11] X. Qiu, M. Marek-Sadowska and W. Maly, "3D chips can be cool: thermal study of VeSFET-based ICs", Proc. IEEE 63rd Electronic Components and Technology Conf., pp. 2349-2355, 2013, DOI**:** 10.1109/ECTC.2013.6575912

[12] H. Wu,W. Wu, M. Si and P. D. Ye, "First demonstration of Ge nanowire CMOS circuits: lowest SS of 64 mV/dec, highest gmax of 1057 μS/μm in Ge nFETs and highest maximum voltage gain of 54 V/V in Ge CMOS inverters", IEDM 2015, DOI: 10.1109/IEDM.2015.7409610

[13] G. Alymov *et al.* "Abrupt current switching in graphene bilayer tunnel transistors enabled by van Hove singularities", Sci. Rep. **6**, 24654, 2016, DOI: 10.1038/srep24654

[14] S. Wachter *et al*. "A microprocessor based on a two-dimensional semiconductor", Nat. Commun. **8,** 14948, 2017, DOI: 10.1038/ncomms14948

[15] C. Qiu *et al.*, "Scaling carbon nanotube complementary transistors to 5-nm gate lengths", Science, vol. 355, pp. 271-276, 20.01.2017, DOI: 10.1126/science.aaj1628