Abstract—In this paper, the opportunities, limits and challenges for future silicon-based field effect transistors (FETs) tailored for radio frequency (RF) and millimeter (mm)-wave circuit design are discussed. After the review of CMOS FET scaling down to the 10 nm node, advanced CMOS techniques such as silicon on insulator (SOI), strained silicon, high-k, low temperature and multigate transistors are treated. Moreover, emerging Beyond CMOS FET concepts based on silicon nanowires, graphene and carbon nano tubes (CNTs) are discussed as potential replacement or extension to CMOS.

Keywords: CMOS scaling, SOI, strained silicon, multigate, FinFET, silicon nanowires, CNT, graphene, Beyond Moore.

I. INTRODUCTION

In 2003, the IBM Microelectronics CTO Dr. Bernard S. Meyerson, stated at the International Electronics Forum in Prague: “Scaling is already dead, but nobody noticed it had stopped breathing and its lips had turned blue”. Indeed, especially for RF circuit design, scaling alone is not sufficient anymore to boost performance from the technology side: worse yet, scaling may even induce major drawbacks. Advanced CMOS concepts such as SOI, strained silicon, high-k, low temperature and multigate transistors can be applied to mitigate some of the problems associated with scaling. However, from a long-term perspective, also these approaches will not be able to prevent CMOS scaling from running into physical limits, which are supposed to be around the 5 nm node [1]. In this regard, Beyond CMOS concepts based on silicon nanowires, carbon nanotubes (CNTs) and graphene are promising alternatives [2]. This paper reviews those approaches and is organised as follows: in Sections II and III, we will review important FET RF performance parameters and the limits of CMOS scaling; advanced CMOS concepts are discussed in Section IV; Beyond CMOS concepts are treated in Sections V; Section VI concludes the paper.

II. REVIEW OF FET RF PERFORMANCE PARAMETERS

The most commonly used speed performance parameter of transistors is the transit frequency \( f_t \), which is the frequency where the extrapolated small-signal current gain

\[
|h_{21}(f)| = \left| \frac{i_g}{i_d} \right|_{v_{ds}=0} \quad \text{equals unity.}
\]

The gate current, drain current, drain source voltage and frequency are denoted by \( i_g, i_d, v_{ds} \) and \( f \), respectively. Here, per definition \( v_{ds} \) has to be zero, which corresponds to a small-signal short-circuit at the transistor output. Hence, the impact of transistor parasitics parallel to the output and in series to the input are ignored. This definition may be sufficient as a first-order performance evaluation as long as the transistors exhibit a low drain source small-signal output conductance \( g_{ds} \) in parallel to the load admittance \( g_L \). However, as discussed in the next sections, for aggressively scaled CMOS circuits as well as for many “Beyond CMOS” approaches, typical \( g_{ds} \) values are so high that, regardless of the load \( g_L \), no small-signal voltage gain can be achieved anymore. The maximum possible voltage gain in a circuit is limited by the high intrinsic voltage gain \( g_m/g_{ds} \). This is incompatible with many high-speed circuit architectures, whose functionality relies on voltage gain. One example is the cascode circuit, where the current source property (increased output impedance) is in first order improved by the intrinsic voltage gain. Hence, using only \( f_t \) as benchmarking parameter becomes increasingly meaningless and is misleading in the context of RF and analogue circuit design, since it does, e.g., neglect any impractically large \( g_{ds} \). Hence, in addition to \( f_t \), at least the intrinsic gain \( g_m/g_{ds} \) should be added as a parameter characterizing a technology. An even more useful performance parameter is the maximum frequency of oscillation \( f_{max} \), since it includes all transistor parasitics and no impractical terminations are used. It is defined as the frequency where the extrapolated maximum available power gain equals unity. Further key parameters are the linearity and the noise figure.

III. REVIEW OF CMOS SCALING

Fig. 1 shows typical key parameters of CMOS generations from 350 nm (state-of-the-art 1997) to 10 nm (state-of-the-art expected around 2020). Until the 350 nm node we can roughly apply the following scaling: \( f_t = g_m/2\pi C_{gg} \sim l_g^{-1} \), where \( g_m \) is the transconductance, \( C_{gg} \) is the total capacitance connected to the gate and \( l_g \) is the gate length. This means that a reduction of \( l_g \) by a factor of two increases the \( f_t \) by a factor of four. Due to short channel and saturation effects such as, e.g., velocity saturation at high electrical fields in short channels, the scaling of \( f_t \) by means of \( l_g \) has degraded to \( l_g^{-1} \) around the 45 nm node. Thus, scaling from 90 nm to 45 nm results only in a doubling of \( f_t \). Towards the 10 nm node, the impact of scaling on \( f_t \) will be even less pronounced with \( l_g^{-1/8} \) and \( 0 < x < 1 \).

To avoid excessive electrical fields in the channel, \( V_{ds} \) has to be decreased with \( l_g \) leading to a \( V_{ds} \) around 0.5 V for the
10 nm node. In turn this reduces the maximum possible RF power drastically, e.g. by a factor of more than 40 compared to 250 nm technology. Moreover, since the threshold voltage $V_{th}$ cannot be scaled proportionally, the margin $V_{ds} - V_{th}$ is significantly reduced challenging the driving of the next stage.

Fig. 1. Key transistor parameters versus CMOS generations.

For RF circuits, it is mandatory to generate gain. Unfortunately, from the 250 nm to the 45 nm node, the intrinsic voltage gain $g_m/g_{ds}$ has decreased from 15 to 5 due to an increased $g_{ds}$. If this trend continues, extrapolation for 10 nm technology leads to an intrinsic gain close to unity. In this case, it would not be possible to design any useful RF amplifier.

A further point is very critical from a fabrication point of view. An oxide thickness $t_{ox}$ of around 1 nm is required for the 10 nm node to allow sufficient channel control via the gate. Since 1 nm corresponds to just a few atomic layers further reduction of $t_{ox}$ is difficult from a gate leakage, noise, and process variation point of view.

Regarding the impact of scaling on passive components and corresponding drawbacks, the reader is referred to [3].

IV. ADVANCED CMOS

Several techniques can be used to mitigate the drawbacks of CMOS scaling discussed in Section III.

A. Strained Silicon

Strain is one method to enhance the carrier transport properties in transistors. If the active silicon region is placed on a substrate layer with a larger lattice constant, e.g. on silicon germanium (SiGe), strain is induced as illustrated in Fig. 2. This modifies the band-structure within the active silicon region. This modification leads to a lower scattering probability and thus, to a higher mobility in the channel. By using silicon with 20 % Ge portion, the electron mobility has been enhanced by 70 % leading to a speed improvement around 30 % [4].

B. Silicon on Insulator (SOI) Technology

A further option for enhanced $l_g$ scaling is SOI technology. An SOI isolation layer separates the thin active device layer and the main substrate. Thus, the resistivity of the active layer determining $V_{th}$ and the resistivity of the main substrate can be chosen independently. Consequently, relatively high resistivities can be chosen for the bulk substrate without degrading $V_{th}$. This results in reduced parasitic and leakage currents in the substrate. Improvements of $f_t$ in the order of 30% are typically achieved with SOI but at the expense of severe self-heating and increased reliability issues. This technique is also very beneficial for passives typically having a strong coupling to the lossy substrate because of their large area.

Fig. 2. Illustration of straining of silicon by means of silicon germanium.

C. Multigate Transistor

A further promising candidate capable of enhancing the speed and the scaling properties is the DG (Double Gate) FET.

Fig. 3. Illustration of multi-gate MOSFETs [5].
Fig. 3a shows a simplified illustration of a DG MOSFET with aligned planar gates. Two connected gates, one at the top and one at the bottom of the channel are implemented. This leads to several advantages. A better channel control is possible since fields act from both sides. Consequently, more aggressive scaling is possible. In principle, two channels are connected in parallel. Hence, more current can flow at given gate source voltage. The fabrication of DG MOSFETs is challenging. In the horizontally aligned double gate MOSFET, the fabrication and low-loss contacting of the back gate is difficult. Both gates should preferably provide the same delays to modulate the channel in phase. Full symmetry can be obtained by applying laterally aligned gates. Corresponding devices can be realised as DG FinFETs or vertical DG FETs as depicted in Fig. 3b, c, respectively. Lower gds and higher gm/gds are possible due to the full depletion of the fins [6]. Further advancements should be possible by applying tri-gate [7] or ring-gate FETs as sketched in Fig. 3d and 3e. However, for the ring gate structure, cost 3-D processing is required.

D. High Barrier Gate Oxide

Gate tunnelling through the very thin oxide layers is a major limitation for CMOS scaling. According to [7], this leads to gate leakage currents that increase exponentially as tox is scaled down. The leakage generates additional noise in circuits. Moreover, in VLSI circuits, leakage leads to a huge increase in overall power consumption. Typically, FETs with 20 nm tox have leakage current densities in the order of 10^{-2} - 10^{-1} A/cm². One possibility to mitigate leakage is to use materials with a high dielectric constant. This allows to increase the oxide thickness while ensuring the same gate control. The thicker the gate oxide, the lower is the gate tunnelling current. Promising candidates of these high-k materials are e.g. zirconium oxide (ZrO2) and hafnium oxide (HfO2). The latter decreases the gate leakage current by a factor of more than 10⁸ for a 20-nm device [8]. However, the manufacturing of such high-k dielectrics is challenging since impurities and traps at the interfaces can degrade RF performance and thus have to be avoided.

E. Low Temperature

The potential to improve the performance of MOSFETs at low temperatures has been known for a long time. Cooling systems for servers have been developed and have been on the market for a decade. Major reasons for the enhancements are higher carrier mobility and lower interconnect resistance yielding performance gains in the order of 1.5 to 2 compared to the operation at room temperature. Nevertheless, cooling has considerable disadvantages. Mandating additional supply power and significantly increased system size, the approach is typically not suited for mobile and low-power applications.

V. BEYOND CMOS APPROACHES

Several techniques have been discussed in Section IV, which are beneficial to delay the CMOS scaling limits at least to the 10 nm node. But what is next once the CMOS limits will be reached? Since scaling alone is not sufficient anymore novel approaches have to be found. Three of the most prominent emerging FET approaches, which are based on graphene, CNTs and nanowires, are discussed in the next sections. These transistor concepts have the potential for better channel control and reduced short channel effects compared to CMOS. Therefore, they allow more aggressive scaling. It is worth mentioning that these transistors can also provide ambipolar behaviour, which is beneficial for designing reconfigurable logic gates. At least within a portion of the transfer characteristics and especially for the CNTs, these devices can show a linear relation between drain current and gate-source voltage. Hence, the level of higher order harmonics and intermodulation products is much smaller. This opens up new opportunities for improved linearity in circuits operating below their compression.

A. Graphene

Referring to Fig. 4, graphene is a 2-D material consisting of a mono-atomic layer of carbon atoms ordered in a honeycomb structure. In [9], fT and fmax values as high as 100 GHz and 10 GHz were published for devices with 240 nm tox. Recently, an fT even above 200 GHz [10] was reported. However, from perspective of circuit design, these numbers are not useful since they are extrapolated for the intrinsic material and exclude major parasitics, such as those of the gate, drain and source fingers and metal interconnects, which have to be considered for realization of real circuits. Hence, the actual fT and fmax values are significantly lower. Some major drawbacks make the design of RF circuits in graphene technologies challenging. In [9], no current saturation is observed up to 2V Vds or before breakdown. Thus, the gds value is very high and consequently, the intrinsic voltage gain is lower than 0.4 for this RF technology. Therefore, typical circuit concepts requiring voltage gain cannot be applied. In addition, it is technologically challenging to fabricate graphene nanoribbons (GNRs) with a non-vanishing bandgap. Thus, the on/off current ratio and the performance of switches are in general very poor. Nonetheless, a very interesting mixer circuit operating a 10 MHz was presented in [11]. Due to symmetric ambipolar conduction, odd-order intermodulation products can be effectively and easily suppressed leading to lower spurious emissions. However, despite the very high measured mobilities in suspended graphene [12] looking very promising for the first view, the mobility degrades significantly after gate oxidation deposition. In GNR transistors, the mobility is additionally reduced by line-edge roughness. Thus, the achievable mobilities can fall below the value typical for conventional semiconductors [13].

B. Carbon Nanotube

A CNT can be viewed as a rolled-up graphene sheet with a diameter of 1 to a few nm. FETs based on CNTs are promising
candidates for RF circuits. Due to the 1-D transport on the surface, not only excellent channel control is achievable, but more importantly, a high linearity is expected due to the linear dependence between drain current and gate source voltage above threshold [14]. Due to the large mean free path, CNTFETs have the potential for low thermal noise and the operation up to THz frequencies [15]. Recent publications [16] show for \( \text{i}_g = 450 \) nm a measured \((f_t, f_{\text{max}})\) up to \((10, 12)\) GHz (only pads de-embedded). In [17], \( f_t \) up to 80 GHz is claimed for a device with 300 nm \( \text{l}_g \). However, such high values are obtained after de-embedding of not only the pads (as in conventional technologies) but the complete metallization. Hence, this value is not directly relevant for real circuits.

In addition to their advantages, the RF impairments of CNT technology have to be considered. Since one CNT tube can carry only a current in the 10 - 30 \( \mu \)A range, arrays of hundreds of CNTs are required to obtain drain currents in the mA range [18], as required for typical RF circuits. Unfortunately, the fabrication techniques known up to date do not allow growing semiconducting CNTs only. In practice, around 30% of the CNTs are metallic. This leads to a poor \( \text{g}_{\text{ds}} \) and consequently a low intrinsic gain. In [17], a \( \text{g}_{\text{m}}/\text{g}_{\text{ds}} \) of only 0.03 to 0.4 was reported.

Referring to Fig. 5, an RF amplifier circuit with 11 dB gain at 1.3 GHz was demonstrated in [19]. Here, \( \text{g}_{\text{m}}/\text{g}_{\text{ds}} \) is above unity. A broadcast radio operating around 1 MHz was presented in [20]. Both circuits are not fully integrated but are assembled with packaged CNTFETs and external passives.

We have designed a 2.4 GHz CNTFET down conversion mixer with 0 dB conversion gain [21]. These results are based on simulations using a reliable CNT model.

Ambipolar CNTFETs exhibit one very interesting property: by means of an additional gate terminal it is possible to change one hard-wired transistor from NMOS to PMOS behaviour, and vice versa. This enables novel logic gate circuits with more functions and, consequently, more computing complexity per transistor. In [22] it was shown theoretically that a reconfigurable circuit with 8 basic logic functions is feasible using only 7 ambipolar transistors.

C. Silicon Nanowire

Silicon nanowire FETs may be a further interesting alternative for RF applications. However, until today most silicon nanowire FETs are optimized for dc and digital applications. In this case, typically, a back gate structure as illustrated in Fig. 6a is used, which allows a relatively simple fabrication process. Small nanowire diameters are favourable to exploit quasi 1-D properties. Since the drain current in a speed-optimized nanowire is small, it is difficult to measure the RF performance of silicon nanowire transistors. A further practical challenge is to realize transistors with arrays of nanowires yielding sufficient current. A sophisticated structure well suited for RF applications is shown in Fig. 6b. It is similar to the CMOS architecture drafted in Fig 3e and is very difficult to fabricate today. Simulations presented in [23] for structures with 30 nm \( \text{l}_g \) indicate promising performances with \( f_t \) up to 500 GHz, \( f_{\text{max}} \) up to 800 GHz and a \( \text{g}_{\text{m}}/\text{g}_{\text{ds}} \) of up to 3. But in these simulations the extrinsic parasitics of the metallisation are not included. The ring gate concept may also be very interesting for CNTFETs.

D. Comparison

A comparison of the key characteristics of high speed graphene, CNT and nanowire FETs is given in Table I.
operation up to THz frequencies and individual advantages, e.g. ambipolarity, and the high linearity of CNT devices. However, all approaches currently face significant technological challenges. Since they have no saturation region, graphene FETs have a large $g_{ds}$. The latter has to be reduced to improve their poor intrinsic voltage gain to values well above unity. Since graphene material has no intrinsic bandgap, graphene FETs have a very low on/off current ratio.

Methods have to be found to reliably fabricate arrays with several hundreds of CNTFETs to yield sufficient drain current. Here, the number of parasitic metallic tubes has to be reduced to decrease the high $g_{ds}$.

For the mentioned recent CNT and graphene FET technologies no voltage gain can be achieved strongly limiting the circuit design flexibility. Here, power gain is based on current gain rather than voltage gain.

To exploit the full speed and scaling potential of silicon nanowire FETs, methods have to be found to realize reliable devices with arrays yielding enough current. Since they have a saturation region and no metallic tube problem, they yield the best $g_{ds}/g_{ds}$ allowing voltage gain in circuits.

The Beyond Moore technologies have to be improved accordingly, with tight interaction to circuit designers.

<table>
<thead>
<tr>
<th>Properties</th>
<th>Graphene</th>
<th>CNT</th>
<th>Nanowire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport</td>
<td>2-D</td>
<td>1-D, quantum effects</td>
<td>Quasi 1-D for very small diameters, otherwise 3-D</td>
</tr>
<tr>
<td>Potential for high linearity due to $I_d$ – $V_{ds}$</td>
<td>Requires very narrow device width (few nm)</td>
<td>Yes</td>
<td>If device diameter is very small</td>
</tr>
<tr>
<td>Scalability of drain current</td>
<td>By device width (parallel nano ribbons)</td>
<td>By number of tubes in parallel</td>
<td>By diameter and parallel wires</td>
</tr>
<tr>
<td>Bandgap</td>
<td>Only for nano ribbons and still very small</td>
<td>Yes for semiconducting tubes</td>
<td>Yes, depending on material</td>
</tr>
<tr>
<td>On/off ratio</td>
<td>Poor since no or small bandgap</td>
<td>Currently poor due to metallic tubes</td>
<td>Good</td>
</tr>
<tr>
<td>Ambipolar properties</td>
<td>Possible</td>
<td>Possible</td>
<td>Possible</td>
</tr>
<tr>
<td>$I_{dr}/I_{tr}$ intrinsic at $V_g$</td>
<td>n.a./100GHz @ 240nm, measured [9]</td>
<td>15GHz/80GHz @ 300nm, top gate, meas. [17]</td>
<td>n.a./up to 500GHz @ 30nm with ring gate, simulation only [23]</td>
</tr>
<tr>
<td>$g_{ds}$ at $V_g$</td>
<td>10GHz @ 240nm, measured [9]</td>
<td>3GHz @ 300nm, top gate, measured [17]</td>
<td>Up to 800GHz/n.a. @ 30nm with ring gate, simulation only [23]</td>
</tr>
<tr>
<td>Intrinsic voltage gain $g_{ds}$</td>
<td>$&lt; 0.4 @ 240nm$, measured [9]</td>
<td>$&lt; 0.4 @ 300nm$, top gate, measured [17]</td>
<td>Up to 3 @ 30nm with ring gate, simulation only [23]</td>
</tr>
<tr>
<td>Challenges for RFIC design using recent technologies indicating demands for technology improvement in future</td>
<td>Very high losses due to large $g_{ds}$, since no current saturation, no voltage gain possible with devices reported in [9] Poor switches since no bandgap Large process tolerances due to edge effects</td>
<td>High losses due to parasitic metallic tubes and associated high $g_{ds}$, no voltage gain possible with [16, 17] Poor switches due to metallic CNTs Fabrication of large number of parallel CNTs to obtain currents in mA range</td>
<td>Fabrication of large number of parallel wires required to obtain currents in mA range Large process tolerances due to edge effects Voltage gain possible with [23]</td>
</tr>
</tbody>
</table>

REFERENCES