Chapter 2
Silicon Technologies to Address mm-Wave Solutions

Andreia Cathelin and John J. Pekarik

2.1 Why Silicon?

There are strong reasons not to consider silicon technologies for mm-wave applications. Silicon comes up short in many comparisons to III-V semiconductors. Silicon carrier mobility is relatively low and so device-level FOMs of raw performance appear to be inferior. The silicon bandgap is relatively small and so voltage tolerance tends to be lower. Furthermore, highly-resistive or semi-insulating silicon substrates are difficult to achieve resulting in poorer isolation and higher losses in interconnects and passive devices. Each of these presents serious challenges to implementing mm-wave functions.

However, advances in silicon technology driven by high-performance digital applications, offer advantages to the mm-wave designer that might not be apparent on first consideration. Performance, quantified by $f_T$, $f_{max}$ or $NF_{min}$ for example, has dramatically increased with geometry scaling and technology enhancements in both CMOS and SiGe HBTs [1]. Both CMOS and BiCMOS technologies have been used to demonstrate circuit functioning at frequencies in and above the K-band. Now, these silicon technologies are, by virtue of nanometer-scale design rules, able to implement staggering amounts of digital logic in a given area thereby enabling the on-chip integration of sophisticated control logic for performance tuning and/or digital signal processing. Furthermore, the worldwide manufacturing capacity of silicon technologies driven by consumer applications like gaming and personal electronic appliances assures low-cost. This will certainly provide an impetus for the evolution of mm-wave consumer applications. The combination of mm-scale wavelengths, low cost and the ability to integrate begs the consideration of array-based transceiver topologies being implemented on a single die or package.

Andreia Cathelin
STMicroelectronics, Crolles, France e-mail: andreia.cathelin@st.com

John J. Pekarik
IBM Corporation, Essex Junction, Vermont, USA e-mail: pekarik@us.ibm.com
2.1.1 Performance

The International Technology Roadmap for Semiconductors (ITRS) [1] hopes to “ensure cost-effective advancements in the performance of the integrated circuit and the products that employ such devices” by compiling and publishing roadmaps which “identify critical challenges” and “encourage innovative solutions” with “focus on technologies that are crucial to the design, materials, and manufacture of semiconductors, as well as the factory sciences, process control, metrology, and environmental aspects.” One chapter of the ITRS focuses on Radio frequency and analog/mixed-signal (RF and AMS) technologies for wireless communications. In tables therein, the demands of a wide range of applications (cellular phones, wireless local area networks, wireless personal area networks (PAN), phased array RF systems, and other emerging wireless communication, radar, and imaging applications) operating between 0.8 GHz and 100 GHz are used to define needed capabilities of silicon and III-V compound semiconductor devices.

Fig. 2.1 shows the roadmap of the cutoff frequency ($f_T$) comparing a number of III-V semiconductor devices with the silicon CMOS NFET and SiGe HBT as taken from the 2006 ITRS [2]. Inasmuch as $f_T$ is an adequate measure of transistor performance, the ITRS shows that silicon transistors are at least competitive with transistors made in III-V semiconductors. Furthermore, the RF & AMS chapter of the ITRS tabulates the $f_T$ of the NFET from the Low Standby Power digital CMOS roadmap whereas the NFET from the High Performance digital CMOS roadmap is shown as having an intrinsic switching speed a factor of three higher. We have noted some of the challenges inherent with the use of silicon transistors and will discuss more below. However, it is evident that silicon technology currently exhibits small-signal gains that are competitive with those of III-V transistors and are predicted to scale at least as quickly in the near-term future.
2.1.2 Cost, Integration [3]

If silicon technology has adequate performance to implement the front-end portions of the transceiver, the ability to integrate digital logic in CMOS at increasing densities offers the opportunity to drastically lower overall system cost. Lower cost could be the prime motivator for the use of BiCMOS or CMOS over III-V technologies. Again considering $f_T$ as a measure of performance, the SiGe BiCMOS HBT has comparable performance to the NFET at roughly twice the minimum feature size. For stand-alone RF functions, where area is dominated by passive devices and I/O pads, BiCMOS may be the lower-cost option despite the approximately 20% additional process complexity required to form the HBT. III-V transistor performance at substantially relaxed lithography dimensions is comparable with leading edge CMOS. So, again for purely RF devices, III-V implementations may be lower cost especially when utilizing existing designs and time-to-market is considered. However, when even modest amounts of digital logic are to be integrated, CMOS has a clear advantage as circuit density and chip size scale with the square of the minimum lithographic dimension.

This can be illustrated by comparing integration densities of BiCMOS and CMOS at equivalent $f_T$. The square of the second metal wiring level (M2) pitch is a measure of the integration density of the digital CMOS available at each technology node as illustrated in Fig. 2.2. Any digital content comes at that node’s CMOS density (and performance). For a hypothetical chip that is 50% digital CMOS and 50% RF implemented in 0.35 μm BiCMOS that is being migrated to 0.18 μm CMOS, if we assume that the RF portion, being dominated by passive devices, does not scale, we see that the overall chip area shrinks by ∼25% due to the dramatic increase in the density of the digital circuitry. By fitting proportionally more chips on a wafer that costs about the same, the economic benefit is obvious. Such an analysis needs to consider all aspects of product cost such as masks, packaging and volume discounts and needs to use current pricing information as competitive pressures usually force prices lower with time. Furthermore, many other considerations will factor into a decision on which technology to use for a given product IC. Designer expertise or the existence of verified circuit designs could drastically affect the time to market. Package, board and system assembly costs could well dominate the contribution of the IC die to the total product cost.

2.1.3 Manufacturing Capacity

The worldwide silicon manufacturing capacity for semiconductor wafers is staggering. It has increased by about 67% from the beginning of 2001 through the first half of 2007, and during that time enough wafer area was manufactured to cover roughly a fourth of the land area of Manhattan. Fig. 2.3 shows data from Semiconductor International Capacity Statistics (Sicas) [4]. The graph shows MOS capacity which includes BiCMOS. Bipolar-only capacity is not shown but that represents less than
5% of the total silicon area being produced. It is evident that the growth is almost exclusively in the smallest lithography node. Also, the capacity of older nodes tends to stabilize to a value slightly below its peak and remain on line for a long time. These two observations are interesting in that most of the demonstrated millimeter-wave circuits have been in the 130nm and smaller nodes. The capacity at 130nm appears to have stabilized and we can now expect commodity pricing pressures to come to bear. Anticipating historical trends will be repeated, the next major node at 90nm may have already reached peak capacity with rapid grow occurring at 65nm and 45nm capacity beginning to emerge. Given the added transistor performance at these smaller nodes, over half of the worldwide silicon manufacturing capacity is ready for the production of millimeter-wave devices.

### 2.2 Modern SiGe and CMOS Technology

#### 2.2.1 Lithography

The performance trend illustrated in Fig. 2.1 is the continuation of a three-decade long trend primarily driven by advances in optical lithography. As illustrated in Fig. 2.4, the wavelength used in lithography has evolved to shorter and shorter lengths but, for some time now, on-wafer dimensions have been below the wavelength of light used to print them. A multitude of techniques are employed to allow printing sub-
wavelength features. Design rules evolved from restrictions on minimum allowed linewidth, space and pitch to restrictions on pitch and orientation with further restrictions on feature placement and shape being considered. On masks, the use of optical proximity corrections (OPC) and sub-resolution assist features (SRAF) aids in the creation of as-printed patterns that more closely resemble as-design shapes. For critical mask levels, the gate, contacts and first metal, the use of phase-shifting is typically employed. The increasing complexity of calculating the optimum deployment of these on-mask techniques is driving the field of computational lithography [5].

Photoresist layer thicknesses have necessarily shrunk with the decrease in feature size. Thinner resist layers are unable to withstand etches or block ion implants to the extent of thicker resists. To aid in these functions, hardmask layers, for example silicon dioxide which selectively withstands silicon etches, are introduced into the process as a necessary complication. Softmasks, non photo-active polymer films, are used when additional resist thickness is needed and are sometimes included as anti-reflection coatings (ARC) to reduce optical interference effects. Photo resist can also be etched after images are developed to achieve smaller printed images, albeit without pitch reduction. Advances in lithography tooling include the use of off-axis illumination to improve depth of focus and automatic dose adjustment either by feedback from prior-lot measurements or in-situ measurements [2]. Major lithography technology advancements include immersion lithography now being introduced into production [6] and extreme ultraviolet (EUV) lithography still in development [7].

The dimensional shrinking of the physical structures on the wafer has led to a series of dramatic changes in materials and structures used in aggressively scaled CMOS process technologies. The following sections briefly touch on three of these that are likely to have particular impact to mm-wave design, the introduction of low-K dielectrics in copper wiring levels, the use of strain engineering to enhance carrier mobility, and the use of metal gates and high-K dielectrics in FETs.
2.2.2 Low-K Dielectrics and Copper Wiring

Copper wiring was introduced in the mid-1990’s to reduce RC-delay in narrow signal lines and to improve robustness to electromigration. The addition of low-K dielectric materials provides an additional reduction in wiring delay. A comparison of representative circuit delay, segregating wire and gate delays is shown in Fig. 2.5. A reduction of wiring delay of almost $3.5 \times$ is illustrated and this full benefit is realized at the smallest dimensions where wiring delays dominate transistor delays. Recent utilization of porous low-K material, with relative dielectric constants as low as 2.4, provides additional incremental improvement as illustrated in Fig. 2.6. In order to assure uniformity in wire dimensions both horizontally, limited by lithography, and vertically, limited by etch and chemical-mechanical polishing (CMP), ground rule restrictions on wiring density are established. Furthermore, foundries often introduce perforations in wide metal shapes and introduce small tiles of metal to fill sparse areas. The impact of these is discussed in Section 2.4.

2.2.3 Mobility and Strain Engineering

For CMOS, through the 90nm node, the scaling of gate length and gate oxide thickness was roughly proportional. Further scaling of oxide thickness is limited by tunneling current and therefore, since the transconductance ($g_m$) is inversely proportional to the oxide thickness, any scaling of $g_m$ must be accomplished through an increase in carrier mobility. Mechanical strain, which distorts the semiconductor crystal lattice, also distorts the energy band structure resulting in higher or lower carrier mobility depending on the carrier type (electrons or holes), whether the strain is compressive or tensile, whether the strain is uniaxial or biaxial, and the magnitude of the strain. Biaxial tensile strain in silicon and compressive or tensile biaxial strain in SiGe can
be induced with epitaxial growth on SiGe buffer layers yielding enhanced electron and hole mobilities and enhanced NFET and PFET performance[10]. Uniaxial strain can be induced by depositing stressed films on top of the completed FET prior to contacts [11] and if the stressed film is deposited prior to the final anneals, the stress is preserved in the device structure even after removal of the film [12]. These techniques are very effective in imparting tensile strain in the channel which thereby improves NFET performance. Compressive uniaxial strain to improve PFET performance has been achieved by selective epitaxial growth of SiGe in the source/drain
Fig. 2.7 Increase in FET drive current due to mobility enhancement techniques [14] (© IEEE 2005).

region adjacent to the gate [13]. Combining these techniques can lead to dramatic simultaneous improvement in both NFET and PFET performance [14].

Carrier mobilities are dependent on crystal orientation for example, hole mobility is higher in the (110) direction while electron mobility in higher in the (100) direction. Silicon wafers using hybrid orientation technology (HOT) allow the simultaneous use of preferred crystal directions for NFETs and PFETs with a single gate orientation. Silicon wafers used in CMOS are conventionally oriented with gates in the (100) crystal direction [15]. Orienting the gates in the (110) direction and applying stressor films is another way to achieve simultaneous NFET and PFET improvement [16].

2.2.4 Metal Gates & High-K Dielectrics

In order to avoid the high tunneling currents resulting from thin gate oxide, dielectric materials with higher effective permittivity (high-K) are being introduced. For an equivalent or lower effective oxide thickness these materials can be physically thicker and have dramatically lower tunneling current [17]. The introduction of metal gate electrodes addresses the problem of polysilicon depletion in the high vertical electric fields now inherent in scaled CMOS and the problem of high gate resistance from polysilicon sheet resistance and contact resistivity. Coupling high-K dielectric materials with the appropriate choice of metal gates leads to FETs with high drive current [18] and improved high-frequency performance [19]. These technology elements are to be employed in high-performance 45nm digital CMOS [20].

2.3 Active Devices on Recent Bulk and SOI Technologies

This section gives an overview of several silicon technologies compatible with very high frequency operation: SiGe:C BiCMOS and bulk and thin SOI deep submicron CMOS. It also presents some active and passive device insights very useful for the millimeter-wave designer [21].
2.3.1 Bipolar Devices

Silicon Heterojunction Bipolar transistors offer some advantages compared to CMOS devices such as lower $1/f$ noise, higher output resistance and higher voltage capability for a given speed [22], [16]. The range of technologies on the market today offers HBTs with $f_T > 200$GHz and sometimes $f_{max} > 300$GHz, as depicted in Fig. 2.8 and Fig. 2.9. Different types of E-B structures are presented, but those obtaining $f_T > 200$GHz and $f_{max} > 300$GHz possess fully self aligned architectures (FSA) and a high performance collector. And finally, the $f_{max}$ parameter for an HBT, has reduced sensitivity to layout parasitics when compared to a MOSFET.

Fig. 2.10 and Fig. 2.11 show $f_T$ and $f_{max}$ data gathered from several foundries, together with the very aggressive ITRS road-map (these two figures should be read together, as the given performances correspond to a given technology node). Differences may come from the differentiation between pure bipolar and BiCMOS processes.
Fig. 2.9 $f_{\text{max}}$ - $BV_{\text{CBO}}$ chart built with various Si/SiGeC HBTs available in 130-nm CMOS node. Different architectures with different maturities are compared (© IEEE 2006).

Fig. 2.10 $f_T$ - $BV_{\text{CEO}}$ chart for different technology nodes.

### 2.3.2 CMOS devices

CMOS transistors follow the well-known Moore’s Law of scaling, thus leading to always increasing functional integration. The 65nm node still uses polysilicon gate, but the carrier mobility is sometimes increased by using several technological solutions as described previously. For MOS devices, $f_T \propto \frac{1}{L_g^\alpha}$, where ($\alpha \sim 1$) and, as a first order approximation, is independent of the gate oxide thickness, as depicted in Fig. 2.12. $f_T$ as high as 150GHz and 200GHz are reached in the 65nm node for Low Power (LP) and General Purpose (GP) devices, respectively. Data gathered
from several major semiconductor foundries show good conformity with the ITRS road-map, as depicted in Fig. 2.13.

The following concerns apply to both bulk and SOI devices, as long as they show the same gate length and the same carrier mobility. Equation 2.1 presents one way of calculating the transition frequency $f_T$ [23] for deep submicron technologies:

$$f_T \approx \frac{g_m}{2\pi C_{gin}} \sqrt{1 + 2\frac{C_{Miller}}{C_{gin}}} = \frac{f_c}{\sqrt{1 + 2\frac{C_{Miller}}{C_{gin}}}}$$ (2.1)
Fig. 2.13 $f_T$ ITRS road-map and several foundries’ performances.

With $g_m$, the gate transconductance, and as also depicted in Fig. 2.14

$$C_{gin} = C_{gsi} + C_{overlap} + C_{fringing}$$  \hspace{1cm} (2.2) \\
$$C_{Miller} = C_{gdi} + C_{overlap} + C_{fringing}$$  \hspace{1cm} (2.3)

and the intrinsic cut-off frequency

$$f_c = \frac{g_m}{2\pi C_{gin}}$$  \hspace{1cm} (2.4)

Where:

- $C_{gsi}, C_{gdi}$ = the equivalent capacitance induced by the source / drain field effect into the channel (see Fig. 2.15)
- $C_{overlap}$ = the equivalent capacitance given by the LDD (low doped drain / source regions) diffusion under the gate (see Fig. 2.15)
- $C_{fringing}$ = the parasitic capacitance depending on the gate height and on the contact to gate distance (see Fig. 2.16).

As it can be seen, $f_T$ increases for reduced gate length devices thanks to higher transistor transconductance [24]. Concerning the capacitive part, $f_T$ is degraded by a high gate to drain capacitance (called also the Miller capacitance).

From a theoretical point of view, the transistor’s transconductance is increasing with the reduction of the effective gate length. Nevertheless, for advanced deep submicron technologies this trend is not straightforward. The features limiting this evolutionary trend are the following:
Fig. 2.14 MOS transistor small-signal equivalent schematic for $f_T$ and $f_{max}$ calculations.

Fig. 2.15 Schematic illustration of the Cgdi and Coverlap equivalent capacitances, CMiller for a MOS device.

Fig. 2.16 Illustration of the Cfringing parasitic capacitance, CMiller for a MOS device.
• Reduced gate oxide thickness. This feature induces gate tunneling current, which is a critical point for Low Power technologies.
• Active zone doping increase. This feature generates mobility degradation.
• Low doped drain regions (LDD). This feature increases directly the source and drain equivalent series resistance.

As discuss in Section 2.2, several process techniques have been hence developed in order to cope with these transconductance increase limitations. As presented in [16], the hole mobility in PMOS transistors is increased by using 45° rotated devices and for the electron mobility in NMOS, tensile liner films are used. In order to limit the gate leakage and to further decrease the effective electrical gate length, high-K dielectrics are used as a part of the gate material. High $f_{\text{max}}$ values have also been reported for CMOS devices, such as 200GHz on 65nm LP node. Fig. 2.17 presents $f_{\text{max}}$ evolution with respect to the effective gate length, for several deep submicron technologies coming from different silicon manufacturers.

Equation 2.5 [23] gives one calculation method for the maximum device frequency $f_{\text{max}}$, all the constants having the well-known significance (see also Fig. 2.14). $R_{i}$ is the equivalent non-quasistatic resistance.

$$f_{\text{max}} \approx \frac{g_m}{2\pi C_{\text{gin}}} \frac{1}{\sqrt{(R_g + R_s + R_i) \left(g_d + g_m \frac{C_{\text{Miller}}}{C_{\text{gin}}}ight)}}$$  

$$= \frac{f_c}{2\sqrt{(R_g + R_s + R_i) \left(g_d + g_m \frac{C_{\text{Miller}}}{C_{\text{gin}}}ight)}}$$

The $f_{\text{max}}$ of CMOS devices, which should correlate to the performance of large signal operation blocks such as mixers, oscillators and power amplifiers, is very sensitive to layout parasitics and also to the choice of the transistor’s finger width. An optimal finger width is chosen for the majority of millimeter wave circuits, all
the solutions seem to converge towards values lower than 5 μm, with two sided gate contacts. A study of this trend is illustrated in Fig. 2.18 [24], where the influence of transistor’s layout on the $f_T$ and $f_{max}$ parameters is evaluated on several transistors with the same total equivalent size, in the 130nm CMOS node. This technology trial study has been held on a SOI technology from STMicroelectronics, but it may of course be extrapolated to any CMOS process. The transistor is divided into several identical cells, and each cell is an interdigitied device with a given unitary finger width.

Several transistor layout topologies are used in order to minimize the extrinsic parasitic elements added to the intrinsic transistor core [25], [26]. The $f_{max}$ parameter is extrapolated from very sensitive high frequency $S$ parameter measurements, and may sometimes suffer from the de-embedding technique. The transistor layout structure presented in [26] is depicted in Fig. 2.19. In this case the transistor is divided into two equal parts with a double contacted gate. This allows minimizing parasitic capacitances and gate resistance while using coplanar access [27]. Source contacts are located on the transistor periphery making easier the contact to the coplanar ground plane. Finally, this structure allows an impedance matched access towards and from the active device to the rest of the circuit.

Table 2.1 presents a comparison between the presented transistor layout and a classical in-line transistor structure [27]. The comparison is done for a CMOS 65nm LP bulk transistor with $W_{total} = 60 \mu m$, $L_g = 60nm$ and with $W_{finger} = 1 \mu m$ embedded in a 6 metal layers back-end (last metal layer is thick). Both structures have 2 sided gate contacts. The extrinsic parasitic elements in this table correspond to an electrical model on the top of a core BSIM4 electrical model presented here as to be the intrinsic NMOS transistor model (see also Fig. 2.20). The intrinsic electrical core model comprises the front-end device plus the contacts and Metal 1 layer on the top of it.
Fig. 2.19 Layout of n-MOS transistors to maximize mm-wave performances and to limit discontinuity [6] (© IEEE 2007).

Table 2.1 Comparison of extrinsic small-signal equivalent circuit parameters for an in-line transistor and the structure presented in Fig. 2.19. LP CMOS 65nm transistor with intrinsic $f_T \sim 160$GHz and $f_{max} \sim 200$GHz, $W_{total} = 60 \mu m$, $L_g = 60nm$, $W_{finger} = 1 \mu m$, 6 Metal BE.

<table>
<thead>
<tr>
<th>Extrinsic parasitic elements</th>
<th>In-Line Device</th>
<th>Fig. 2.19 Device</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cgs (fF)</td>
<td>6.4</td>
<td>3.7</td>
</tr>
<tr>
<td>Cgd (fF)</td>
<td>0.3</td>
<td>3.2</td>
</tr>
<tr>
<td>Cds (fF)</td>
<td>5.3</td>
<td>4.4</td>
</tr>
<tr>
<td>Rs (Ω)</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>Rd (Ω)</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Rg (Ω)</td>
<td>4</td>
<td>0.35</td>
</tr>
<tr>
<td>Ls (pH)</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Ld (pH)</td>
<td>4.5</td>
<td>2</td>
</tr>
<tr>
<td>Lg (pH)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Extrinsic $f_T$ (GHz)</td>
<td>118</td>
<td>138</td>
</tr>
<tr>
<td>Extrinsic $f_{max}$ (GHz)</td>
<td>125</td>
<td>194</td>
</tr>
</tbody>
</table>

The noise parameter $NF_{min}$ is also an important marker for the RF and mm-wave performances of an active device. The very low noise figure values, presented by deep submicron CMOS devices nowadays, are very difficult to measure with an excellent accuracy, as the de-embedding is very complex in the high frequency range. For the MOS transistor, the following formula can be used to calculate the $NF_{min}$ parameter (see also Fig. 2.14).

$$NF_{min} \approx 1 + \frac{f}{f_T} \sqrt{g_m(R_g + R_s)}$$  \hspace{1cm} (2.7)
As it can be easily recognized, this parameter is strongly degraded by high gate and source parasitic resistances, thus being layout dependent. Fig. 2.21 presents some $NF_{\min}$ data, gathered from two silicon foundries and presented with respect to the ITRS road-map. One may observe that similar RF noise behavior may be obtained from different technology nodes transistors, depending on the transistor’s architecture. Nevertheless, both bulk and SOI processes from the same technology node perform the same noise behavior, given that the transistors show the same transconductance and the same parasitic resistances.

Figures 2.22 and 2.23 present the frequency evolution of $NF_{\min}$ for several total widths of a given NMOS transistor ($W$ in $\mu$m), data coming from bulk and SOI FB 65nm LP CMOS devices from STMicroelectronics [28].

### 2.3.3 SOI CMOS devices

As seen from the previous section, moving from bulk to thin SOI does not change the HF performances of CMOS devices significantly [24].

Two kinds of CMOS transistors are available on SOI [29], floating body (FB) and body contacted (BC) devices (see Fig. 2.24). FB devices show the same layout as a bulk CMOS transistor. Thin SOI Partially Depleted (PD) technology induces a floating body effect in the transistor’s channel, which is beneficial for digital operation. Nevertheless, this device suffers from Kink effect (see Fig. 2.25) that degrades
Fig. 2.21 $NF_{min}$ ITRS road-map and several foundries’ performances.

Fig. 2.22 Evolution of $NF_{min}$ with frequency; bulk and SOI 65nm LP CMOS; Lg=60nm, 2 gate access, Wfinger=1 μm, Ids=150 μA/μm, Vds=1.2V (data from STMicroelectronics).

both voltage gain (high $g_{ds}$, see Fig. 2.26) and $1/f$ noise (see Fig. 2.27). This excess low frequency noise occurs only in the Kink region and only for the floating body transistors. The body contact efficiency depends on the body resistance restricting access to the channel, thus a maximum finger width limit should be applied for BC devices to maintain low body resistance. Nevertheless, the low frequency noise behavior of a BC device is identical to the one of a similar bulk transistor. Given all this information, it is obvious that BC transistors are well-suited for the analog part of an RF/mm-wave design.

However, the Kink effect has no impact on HF characteristics and SOI FB MOSFETs exhibit $f_T$ and $f_{max}$ at least identical to those of bulk devices with identical
Fig. 2.23 Evolution of $NF_{min}$ @ 24GHz with $I_{ds}$; bulk and SOI 65nm LP CMOS (data from STMicroelectronics).

Fig. 2.24 Layout difference between Floating Body (FB) - left, and Body Contacted (BC) - right, devices in SOI technologies (130nm technology node).

layout. On the contrary, HF performance of BC devices is penalized by their specific layout (higher gate resistance and gate to source capacitance). The example depicted in Fig. 2.28 shows 65nm LP SOI NMOS transistors that are well suited for mm-wave design:

- for the FB NMOS device: $f_T \sim 160$GHz, $f_{max} \sim 200$GHz
- for the BC NMOS device: $f_T \sim 108$GHz, $f_{max} \sim 126$GHz. In this case this is a worst case value, as because of the particular layout architecture of BC devices, the optimum $W_{finger}$ for those devices is different from the one of FB devices.

As depicted in Fig. 2.21, SOI FB devices show the same noise performance as the bulk ones. Obviously, the BC transistors show larger noise figures, given the increased gate resistance and some capacitive coupling to the body contact. Fig. 2.29 shows measured $NF_{min}$ for FB and BC NMOS devices in a 65nm LP SOI CMOS technology, together with the associated gain.
Fig. 2.25 DC curves for FB and BC devices in PD SOI CMOS.

Fig. 2.26 Comparison of voltage gain for FB and BC devices in 130nm SOI CMOS (data from STMicroelectronics).

Fig. 2.27 Low Frequency noise power spectral density; comparison between FB, BC and bulk transistors in 130nm SOI CMOS (data from STMicroelectronics).
Fig. 2.28  $f_T$ and $f_{max}$ behavior of FB and BC NMOS devices in a 65nm LP SOI CMOS technology (data from STMicroelectronics, IEMN measurement data); $V_{ds}=1.2\text{V}$, $I_{ds}=0.3\text{mA/μm}$, $L_g=60\text{nm}$, $W_{finger}=1\text{μm}$, $W_{total}=64\text{μm}$ (© IEEE 2007).

Fig. 2.29  $NF_{min}$ behavior of FB and BC NMOS devices in a 65nm LP SOI CMOS technology (data from STMicroelectronics, IEMN measurement data); $V_{ds}=1.2\text{V}$, $I_{ds}=0.1\text{mA/μm}$, $L_g=60\text{nm}$, $W_{finger}=1\text{μm}$, $W_{total}=64\text{μm}$ (© IEEE 2007).
2.3.4 Current Density Scaling for CMOS and Bipolar Devices

A very interesting scaling feature has been pointed out in [30], regarding GP bulk and SOI NMOS transistors from different technology nodes. It has been shown that, as a result of constant-field scaling, the peak $f_T$ ($\sim 0.3$ mA/$\mu$m), peak $f_{\text{max}}$ ($\sim 0.2$ mA/$\mu$m) and optimum noise figure $N F_{\text{min}}$ ($\sim 0.15$ mA/$\mu$m) current densities are unchanged from one technology node to another (see Fig. 2.30, Fig. 2.31). This has been proven by different foundries’ measurements with technology nodes from 0.25 $\mu$m down to 90nm. This feature is very convenient when porting CMOS design from one technology node to another. Having in mind these data, constant current-density biasing schematics may be applied to CMOS circuit designs for analog, RF/mm-wave and also high speed digital functions. This way, the impact of statistical process variation, temperature and bias current variation may be drastically reduced.

For HBT devices from different technology nodes, this constant current density for peak $f_T$ is not observed, as depicted in [31] and in Fig. 2.32. Indeed, one efficient way to increase the transistor’s $f_T$ by delaying the onset of the Kirk effect is to increase the collector doping. Hence, the collector current density at peak $f_T$ rises under these circumstances.

2.3.5 Comparison Between State-of-the-Art HBT and CMOS Devices

In general, bipolar devices are specially developed for high frequency / high speed operation modes, while the MOS devices are implemented first for digital applications and then slightly optimized for analog/RF purposes. Fig. 2.33 and Fig. 2.34 attempt to realize a comparison between state of the art contemporary 130nm node SiGe:C HBT and 65nm LP node NMOS devices (data from STMicroelectronics). For this comparison, the dimensions chosen for the two devices are suited geometries for mm-wave design (mostly for low power linear operation). The bipolar device is a 130nm based FSA Si/SiGe:C HBT with 3 emitter fingers of 2.5 $\mu$m each. The NMOS device consists of 10 cells in parallel, each cell of 9 gate fingers of $W_{\text{finger}} = 1$ $\mu$m each and the gates are contacted on both sides.

The first observation is the current difference for the peak value of $f_T$ (see Fig. 2.33). The MOS device has its peak $f_T$ value at a drain current density per unit width of 0.3mA/$\mu$m, while the equivalent value for the bipolar device is reached at a collector current density per emitter length of 1.9mA/$\mu$m. Fig. 2.33 presents also the minimum noise figures measured at 40GHz for both devices. $N F_{\text{min}}$ of 1.1dB and 0.95dB are measured for the HBT and the NMOS devices, respectively. Fig. 2.34 presents the $N F_{\text{min}}$ evolution with the frequency for both devices. Performing very accurate noise figure measurements at such high frequencies with a tolerance lower than 0.1dB is very difficult. When reading the cited figure, for frequencies below 40GHz, the MOS device performs a better noise figure than the
bipolar, but this trend should be reversed for frequencies above 60GHz. The noise figure of the NMOS device can be extrapolated to high frequencies with a good degree of confidence, which would give a value of about 2.7dB at 80GHz. It is to be noticed that $NF_{min}$ is reached for all the frequencies at a constant value of $V_{GS}$. A linear fit can be applied to HBT too, but with a lower level of confidence, as $NF_{min}$ shifts towards a higher $V_{BE}$ when the frequency increases. Nevertheless, the HBT curve in Fig. 2.34 shows a nice fit between measurement and the linear fitted curve, up to 65GHz. An extrapolation using the small-signal equivalent circuit would give $NF_{min} \sim 2.2$dB at 80GHz, and this value is coherent with the one obtained from the linear extrapolation ($\sim 2$dB). As a fair comparison sum-up, it may be admitted that
Fig. 2.32 Evolution trend of the current gain cut-off frequency $f_T$ with the collector current density $J_C$ (technology trials at STMicroelectronics) (© IEEE 2006).

Fig. 2.33 $f_T$, $f_{max}$ and $NF_{min}$ at 40GHz vs. current density for 0.13 $\mu$m BiCMOS and 65-nm LP NMOS (data from STMicroelectronics) (© IEEE 2006).

both the bipolar and the NMOS device exhibit rather similar noise performances. For a given $f_{max}$, the higher current densities and voltage swing, the lower collector to bulk capacitance, along with a higher transconductance, permit the HBT to acquire significant advantage over the MOS transistor for large signal operation blocks such as power amplifiers. Nevertheless, and also to conclude this sub-section, both devices exhibit RF features compatible with millimeter wave circuit integration with industrial margin. The technology choice is given at the end of the day by the type of application, the technology availability and by the SiP or SoC approach chosen for the product.
2.4 Impact of the Back-End of Line on mm-Wave Design

Contrary to what usually happens in analog (and in a given perimeter in RF) design, information on the active device is not sufficient for the millimeter wave designer in order to choose the most appropriate technology. For frequencies above 10GHz, each μm of back-end strip has a significant influence on the electrical behavior of such a circuit. The trend when moving from one CMOS node to the next is the vertical shrink of the BEOL together with the decrease of the metal and dielectric thicknesses and of the metal pitch in order to increase integration density. Diminishing the lateral dimensions (width) of a metallic path imposes also a shrink on the vertical dimensions (height). The generally used rule of thumb in this case gives a shape factor of about three between a metal conductor height and its minimal width [32]. The Back End of Line (BEOL) evolution through technology nodes from 0.35 μm down to 65nm is depicted in Fig. 2.35. One may observe that the general trend is somehow in antinomy with higher and higher working frequencies. In order to insure a very (and ultra) large scale of integration for the active devices, the first metal layers become thinner as do the corresponding Inter Metal Dielectric (IMD) layers. The increase of the integration scale imposes also the use of a larger number of metal levels and also implies some changes in the nature of the metallic materials used. The first technology generations developed for RF design such as 0.35 μm to 0.18 μm had Aluminum layers. Starting from the 0.13 μm generation node, the metallic layers are made of copper which permits, for an equivalent metallic strip conductivity, to diminish the height of the deposited layer.

The connecting line on Fig. 2.35 represents the total dielectric height for a microstrip transmission line built with the last metal layer (for deep submicron technologies, not all the metal layers are supposed to be used as for a digital design). One can observe that the total height of dielectric is diminishing from one technology node to another, thus increasing the influence of the substrate losses on the propagation
The dielectric oxide permittivity tends to diminish from one technology node to the next one, in order to limit the coupling effects between two conducting layers. Nevertheless, the use of thinner layers imposes an increased contribution of silicon nitride (in order to limit the copper diffusion into the oxide), which presents a higher permittivity. With these considerations we can conclude that the total Inter Metal Dielectric height diminishes with the technology nodes and the equivalent dielectric permittivity remains mostly constant.

For localized passive elements such as inductors, the small metal line pitch together with the thin dielectric layers induce larger ohmic and substrate losses. For lumped elements such as transmission lines, the patterned ground shields are not sufficient to limit the attenuation constant degradation. In order to cope with these design difficulties for mm-wave circuits, different trends came out during the latest years. Two families of back-ends may be distinguished: while the main stream is held by the (deep) sub-micron digital processes BEOL, the second trend is presenting millimeter-wave dedicated back-end. Fig. 2.36 illustrates a mm-wave dedicated BEOL for a 130nm node high performance BiCMOS process from STMicroelectronics [31].

It is obvious that, for millimeter-waves dedicated BEOL, the important number of thick metals (3 typically) and the use of thick Inter Metal Dielectrics provides for low losses in the transmission lines. In counterpoint, it imposes the development and the use of an application dedicated process, which sometimes may present higher development costs with respect to standard back end solutions. For a digital BEOL, a large number of metal levels lowers the losses of last metal level transmission lines, but in counterpoint one may look on the economic impact of processes with up to 10 metal levels.

The availability of High Resistivity (HR) SOI substrates for CMOS VLSI integration proposes an alternative to these two solutions, by drastically reducing all
parasitic losses towards the substrate [29], [33] while using standard digital back-end of line. Figure 2.37 presents, as an example, a cross section view on a 65nm CMOS SOI technology from STMicroelectronics. Like other SOI technologies, this one is distinguished by the presence of a buried oxide layer allowing the use of a high resistivity (HR) substrate underneath. The substrate resistivity value is superior to $3\, \kilo\Omega\cdot\text{cm}$. The back end of line is composed of 6 or 7 copper metal layers (depending on the option) and a last encapsulated aluminum layer.

Fig. 2.38 illustrates the impact of the several types of back-end and substrate resistivity on mm-wave transmission lines [34]. Table 2.2 gives detailed explanations on the BEOL and substrate type, as well as implementation data for the compared transmission lines in Fig. 2.38. (Note: HR SOI CMOS measurements have been performed up to 200GHz with a very good confidence level, and gave good concordance with simulated data [33].)

The first comment is that, as expected, the mm-wave dedicated BEOL should allow the lowest attenuation constants for such 50\,\Omega transmission lines, such as
0.5 dB/mm at 60GHz. On the other hand, both coplanar and microstrip transmission lines in a pure digital BEOL from the 65nm node show more losses, such as 2.5 to 3.5 dB/mm at 60GHz. The good surprise comes from the same 65nm node BEOL used this time in a HR substrate SOI technology, where one may observe attenuation constants almost as good as those from the mm-wave dedicated back end from a 130nm node. And finally, there is only 0.1dB/mm of difference between a 130nm and a 65nm BEOL transmission line on HR SOI processes.
Table 2.2  50Ω transmission lines characteristics for several BEOL and substrate resistivities (data from STMicroelectronics).

<table>
<thead>
<tr>
<th>Reference</th>
<th>MS BiCMOS9MW</th>
<th>CPW CMOS130nm HR SOI</th>
<th>CPW CMOS65nm HR SOI</th>
<th>CPW CMOS65nm</th>
<th>MS CMOS65nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>BiCMOS (mmW dedicated):</td>
<td>- 6 metal levels - 3 thick</td>
<td>SOI CMOS 130nm(digital):</td>
<td>- 6 metal levels - 1 thick</td>
<td>SOI CMOS 65nm(digital):</td>
</tr>
<tr>
<td>Substrate</td>
<td>Low Resistivity</td>
<td>High Resistivity</td>
<td>High Resistivity</td>
<td>Low Resistivity</td>
<td>Low Resistivity</td>
</tr>
<tr>
<td>Line type</td>
<td>Microstrip</td>
<td>Coplanar</td>
<td>Coplanar</td>
<td>Coplanar</td>
<td>Microstrip</td>
</tr>
<tr>
<td>Size</td>
<td>W=5 µm (M6+Alu) Distance to ground (M1+M2), h=8.4 µm</td>
<td>W= 26 µm (M1 to Alu) Space to ground (M1 to Alu) s=22 µm</td>
<td>W= 12 µm (M1 to Alu) Space to ground (M1 to Alu) s=5 µm</td>
<td>W=12um (M6 + Alu) Space to ground (M1 to Alu) s=7um</td>
<td>W=3.6um (M6) Distance to ground (M1) h=2.17um</td>
</tr>
</tbody>
</table>

Fig. 2.39  Example of metal layer design rules.

Vertical shrink of the BEOL together with the use of copper layers impose also very drastic layout design rules in terms of metal densities per very tiny areas and also electromigration rules especially for high temperature. The deposition method for copper layers is damascene, based on the principle of dielectric trenches which are filled with metal followed by the CMP step which gives the desired level uniformity. However this uniformity operation is possible only if the underneath metal layers densities are homogeneous all over the wafer. This implies that either full metal drawing or the absence of metals on a large wafer surface is not allowed. Thus, the metallization levels must be perforated if their area is too large and some small tiles of dummy metal must be placed to preserve the homogeneity if the initial density is too low. It should be noticed that these rules of densities follow the projection of the technologies in the future, when these constraints will be more and more severe (smaller control windows). Fig. 2.39 illustrates some of these design rules on one metal level.
Among the rules to be respected for each metal level, let us quote inter alia:

- The minimum and maximum density per a given control window (in general between 20 and 80%)
- The minimum size of an enclosed area in a metallic layer ($W_t$, see Fig. 2.39)
- The minimum and maximum width of a metal stripe ($W_c$, see Fig. 2.39)

All these constraints have to be taken into account at the very beginning of every millimeter wave design. The ground planes contribute a lot to the metal density filling of the total area, but also the gap between the transmission line and these ground planes must respect the metal density rules. The maximum line width is also controlled by DRC rules, thus limiting the theoretical choices for the transmission lines design. On the other hand, the central signal stripe width is determined also by electromigration rules depending on the maximum current density to flow without damage.

An example of such transmission line on a digital BEOL in the 65nm node is presented in Fig. 2.40 and Fig. 2.41 [35]. Classical transmission line theory would have imposed that the ground planes should be built at the same level as the central conducting path, but in the presented case, in order to respect the metal density rules, the ground planes have also a vertical dimension in order to fill the space also for the lower level metals. These extra levels in the ground plane impose extra losses by fringing coupling. 3D electromagnetic field solvers [36], [37] offer good help for these studies.

![Fig. 2.40](image1.png) Coplanar transmission line in a 65nm CMOS digital BEOL (6 Metal layers) (© IEEE 2007).

![Fig. 2.41](image2.png) Top view for the transmission line in Fig. 2.40.
The Back-End of Line of the recent technology nodes permits the integration of passive devices compliant with the design of mm-wave integrated products. A digital back-end on standard VLSI CMOS technologies implies rather lossy solutions for the passives, which have to be overcome by a higher overall current consumption and sometimes figures of merit (such as Noise Figure). In counterpoint, the use of a standard VLSI technology permits a good manufacturing flexibility. On the other hand, mm-wave dedicated back-end permits passive integration with very low losses, thus lower power consumption, but the overall production cost is raised by the use of an application dedicated process. An excellent compromise between these two alternatives may come from the use of VLSI CMOS SOI solutions using High Resistivity substrate, permitting a high integration level with reduced power consumption, for almost the same production cost as standard CMOS.

2.5 Conclusion

Silicon technologies are well-suited for implementing many millimeter wave applications. With the enormous and increasing worldwide silicon manufacturing capacity, reasonably-priced supply is guaranteed assuring the ability to satisfy demand on the consumer scale. Silicon transistor performance, measured by the conventional FOMs, is at least on par with that exhibited by III-V compound semiconductor devices. Furthermore, with the continued, intense focus of the commercial electronics industry, the relentless trend of increasing performance and integration density will doubtlessly continue. Indeed there are challenges to the designer. Design rule stricture and complexity, especially those involving metal pattern density, while certainly inconvenient to digital designers, will prove limiting to conventional millimeter design techniques. The lower voltage tolerance of silicon also presents a serious challenge to conventional design requiring innovation as illustrated by some excellent examples in later chapters. The extent to which the design community can overcome these challenges will determine the degree to which silicon dominates the emerging millimeter wave market.

2.6 Acknowledgements:

The following people are kindly acknowledged for their technical support to this material: from STMicroelectronics: B. Martineau, N. Seller, Ch. Raynaud (also with CEA/LETI), S. Pruvost, D. Gloria, F. Gianesello, P. Scheer, B. Sautreuil, A. Chantre, P. Chevalier, D. Belot, J.P. Schoellkopf and A. Artieri; from IBM: B. Jagannathan and A. Joseph.
References

17. E. Gusev, et al., IBM J. Res. & Dev. vol. 50 No. 4/5 July/September 2006
mm-Wave Silicon Technology
60 GHz and Beyond
Niknejad, A.M.; Hashemi, H. (Eds.)
2008, XV, 302 p. 65 illus., Hardcover
ISBN: 978-0-387-76558-7