Chapter 1
Introduction to mm-Wave Silicon Devices, Circuits, and Systems

Ali M. Niknejad and Hossein Hashemi

1.1 Introduction

Silicon-based RF and microwave technology has had a dramatic impact on the world of wireless technology. Today we can access voice/data and entertainment in virtually every corner of the globe, from short range Bluetooth and WiFi networks, to cellular and satellite networks, to meet different range and throughput requirements. A laptop computer without wireless capability is unthinkable today whereas ten years ago the technologies did not exist at all. What do the next ten years promise? What gaps in wireless technology exist even today? Perhaps the most obvious missing link is between the various peripherals that we carry with us, such as cellular phones or personal digital assistants (or "smart phones" if you prefer), digital cameras, music and video players (such as the ubiquitous iPod), laptops, peripherals such as external hard drives and monitors. The case of the mobile phone is particulary important since the existing wireless connectivity is either too slow and power hungry (Bluetooth) or designed and optimized for longer ranges (WiFi) . What is missing is a wireless USB capability that can support high data rates demanded by large datarate multimedia applications. Wireless technology has been conspicuously absent from MP3 music players (such as Apple's iPod), which are ideal candidates for downloading music and video. While nascent UWB technology is a potential solution, it has some shortcomings including problems with interference and a limited data rate. The 3–5 GHz spectrum is relatively crowded with many interferers appearing in the WiFi bands. A UWB system has very limited transmit power (average of about 0 dBm) which means that bandwidth has to be traded to overcome the limited SNR. In comparison, the 60 GHz band offers the same amount of spectrum (7 GHz), virtually no

Ali M. Niknejad
University of California, Berkeley, e-mail: niknejad@berkeley.edu

Hossein Hashemi
University of Southern California, Los Angeles, CA 90089-0271, e-mail: hosseinh@usc.edu

Fig. 1.1 A hypothetical incarnation of the “iPhone" with mm-wave capability can communicate with several other devices simultaneously using beam forming at a data rate of over 1 Gb/s.

interference[1], and up to 40 dBm of transmit power. Given this bandwidth, it is easy to envision a wireless link capable of supporting multi-Gb/s communication between devices transporting large volumes of media information (video) .

Consider a hypothetical 60 GHz enabled iPhone shown in Fig. 1.1, which uses the 60 GHz connection to download movies from a kiosk (perhaps at a train station or airport), transmits video to a larger screen for easier viewing, and connects to external peripherals such as hard disks and wired and optical networks. If such a device is realized with reasonable power consumption, we see that it can truly displace the laptop computer. Given that this 60 GHz spectrum can be exploited using inexpensive silicon technology, there is quite a lot of excitement and energy in academia and industry focused on silicon mm-wave technology. In Fig. 1.2 we show a plot of the number of papers published worldwide with the keyword “mm-wave", “60 GHz", or “77 GHz". The graph is further separated into papers using CMOS and SiGe. As evident in the figure, interest has grown considerably in this technology in this decade.

Research in mm-wave electronics did not of course start with silicon and predates much of the current generation of work in the area. Active use of mm-wave includes

[1] There is no interference today, but we envision a near future crowded with 60 GHz devices. Nevertheless, the higher propagation loss at 60 GHz means that networks will be naturally spatially isolated.

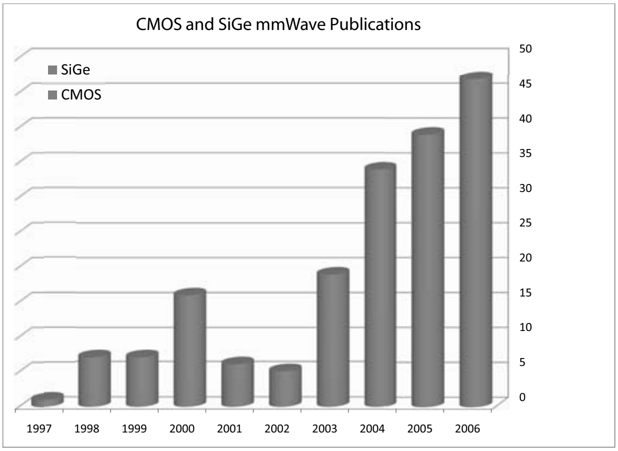

Fig. 1.2 The number of mm-wave papers published in the past decade. Search results obtained from IEEE Xplore database using "60 GHz" or "mm-wave" and "CMOS" or "SiGe" as keywords.

military radar systems, radio astronomy, and space programs. But commercialization efforts, though, have failed to materialize into mm-wave products, mainly due to cost and volume considerations. With the exception of automotive radar in a limited number of luxury vehicles, there are virtually no practical applications for mm-waves today. In fact, when the 60 GHz project was started at the Berkeley Wireless Research Center (BWRC) in 2000, most of our industrial members did not see any potential for the technology in the foreseeable future. Many thought it was good long range research that may materialize in ten years. The pessimists doubted that the technology could be realized with silicon technology, worse yet with CMOS. But in the course of six years, today most of the members of BWRC are actively interested in the developments in 60 GHz and many have begun their own investigations. Many academic institutions worldwide are also actively researching mm-wave technology and many of these efforts are highlighted in this book. Research projects have now explored SiGe and CMOS for mm-wave frequencies beyond 100 GHz.

1.2 Why mm-Waves?

Based on Shannon's Theorem , the maximum data-rate of a communication channel, known as channel capacity, C, is related to the frequency bandwidth of the channel, BW, and the signal-to-noise ratio, SNR, in the following manner [1]:

$$C = BW \cdot \log_2 (1 + SNR) \tag{1.1}$$

which shows that one way to increase the communication data rate is to use more bandwidth. The information bearing signal is usually modulated around a carrier

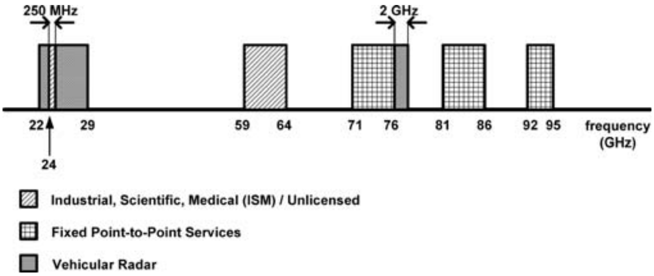

Fig. 1.3 The mm-wave band allocation in the United States.

frequency for proper propagation; therefore more bandwidth is available around higher carrier frequencies. For instance, optical fibers provide a low-loss and wide bandwidth medium, typically tens of GHz, around optical carrier frequencies of 250 THz for high speed wireline communications. The lower part of frequency spectrum is allocated to several narrowband applications such as radio, television, mobile phones, satellite communications, radio astronomy, wireless networking, and various military needs. The Federal Communication Commission (FCC) has allocated several frequency bands at millimeter waves for high data rate wireless communication. Figure 1.3 shows selected parts of the FCC-allocated frequency spectrum. Also shown in this figure are frequency allocations for automotive radar applications. The radar azimuth resolution (perpendicular to the radar wave) and range resolution (in the direction of radar wave) are inversely proportional with the carrier frequency and bandwidth, respectively, explaining the choice of such high frequencies and bandwidths. While the 22-29 GHz frequency band is allocated for short-range applications such as park assist, stop-and-go, and blind spot detection, the 77 GHz band is used for long-range automatic cruise control application.

The second term affecting the communication data rate as well as the radar maximum range is the overall SNR. Unfortunately, for a given distance, the received signal at higher carrier frequencies experiences more attenuation due to the following reasons: First, since the antenna size is normally inversely proportional to the carrier frequency; the higher the carrier frequency, the smaller the antenna size, resulting in less collected power. Higher absorption of air and most other material is the second cause of more signal attenuation at higher frequencies. In a multi-path environment, multiple replicates of the transmitted signal that are reflected from various objects reach the receiver at different times with different amplitudes and phases, causing unwanted signal fading. The amount of attenuation due to unwanted multi-path effects depends on the size of scattering objects relative to the carrier frequency as well as their type and location. A lower SNR reduces the data rate in communication systems for a given distance or reduces the range of wireless communication (e.g., in a radar application). Interference signals also have a noise-like behavior and reduce the received SNR. Fortunately, the larger attenuation at higher frequencies reduces the level of interferences as well as the multi-path components; the latter causes a

smaller delay spread, making it suitable for high speed wireless communication over a shorter range.

Over the years, a variety of modulation, coding, and diversity schemes have been employed to increase the SNR, hence increase the communication data rate. Spatial diversity, consisting of multiple antennas spaced apart, is one of the most attractive methods to increase the SNR. In a statistically fading environment, the received signal varies rapidly as the distance between the transmitter and receiver is altered by only a fraction of a wavelength. Spatially separated antennas can be used to extract more information from these large signal fluctuations, leading to an effectively higher data rate. They can also focus the signal energy into a narrow beam radiating into/from specific directions and place nulls in undesired directions. These beam forming schemes reduce the interference levels, increase the effective SNR, lower the transmitted signal energy to other directions, and are very suitable for radar and imaging applications. Similar to the receive antennas, in a statistically fading channel, transmitted signals from spatially separated transmitting antennas experience independent fades as they reach the receiver. By transmitting different signals from various antennas over time, a larger signal-to-noise ratio at the receiver can be achieved, increasing the overall capacity and performance. In addition to more available bandwidth at higher frequencies that results in higher capacity in data communication or better resolution in radar and imaging systems, the required antenna size and spacing in multiple-antenna systems are also reduced, making the system more compact and affordable.

1.3 The Birth of Silicon mm-Wave

Silicon technology has all but displaced GaAs and other technologies for RF applications in the low GHz regime. A few niche applications, such as power amplifiers, remain as a stronghold but are also under threat by several upstarts. For those with faith in Moore's Law, this was an inevitable consequence in scaling. Transistors became small enough, and consequently fast enough, to operate into the GHz frequencies, thus vying for countless communication applications in these frequency bands.

As the high-frequency capabilities of CMOS improve through scaling, the question is not *if*, but rather *when* will silicon become a viable alternative for mm-wave applications. To answer this question, it is fruitful to consider how silicon-based RF and microwave circuits came about. Even in the $0.25\,\mu\mathrm{m}$ CMOS technology node, researchers and industrial start-up companies demonstrated low-cost radio solutions up to nearly 6 GHz for 802.11a applications. These demonstrations used relatively conventional circuit design and modeling techniques. And yet only ten years later, mm-wave silicon circuits have been widely demonstrated starting with the 130 nm technology node. Clearly the innovations to reach this speed did not come from scaling alone, since the process f_T has only increased by two-fold. Furthermore, increasing the operating frequency by a factor of 10 required a new design

methodology since the wavelength is on the order of the dimensions of the chip components (on-chip $\lambda = 2.5\,\mathrm{mm}$ at 60 GHz).

Silicon was not the obvious choice for 60 GHz systems. Many non-silicon-based technology choices come to mind such as using GaAs MESFETs, PHEMT, InP HEMT, GaAs MHEMT, GaAs HBT, InP HBT. While these exotic technologies offer higher frequency of operation, they are expensive and have low manufacturing yields, and thus offer limited integration possibilities. Furthermore, these processes are not expected to scale in cost as rapidly as silicon (particularly CMOS) technology. Silicon technology enjoys steady scaling in large part due to the investments of industry and government which is tied to a healthy and vibrant multi-billion dollar market for digital, analog, and RF circuits.

Industry's choice to embrace Si rather than traditional III-V technologies was also a philosophical revolution. Engineers trained in the art and science of silicon analog design carried with them a design methodology based on compact models and SPICE-based transient computer simulation. The microwave community, on the other hand, relied on a rich theory based on two-port parameters and extensive measurements to characterize devices. Non-linear analysis was performed with harmonic balance. Active devices were often treated as black boxes with a given N-port response characterized by measurements or simple equivalent circuits. Which design methodology is most appropriate for emerging new applications in the microwave and mm-wave bands? In this book we present both design approaches and highlight where one is particularly more suitable or appropriate.

Economics was the ultimate factor in the decision to use silicon for applications up to 10 GHz. Silicon RF technology is relatively inexpensive, and in the case of CMOS, practically "free." The burden and major cost to develop silicon CMOS technology was provided by a large digital microprocessor market and analog and RF applications could leverage many of the innovations in the manufacturing of silicon circuits. On the other hand, early RF and analog designers using silicon and particulary CMOS were hampered by modeling problems, as the models were often extracted for digital applications. Today the situation is quite different and silicon is the technology of choice for many RF applications, including low-range low-cost radios, short-range high data-rate wireless LAN, cellular radios, and even power amplifiers. SiGe technology has given bipolar technology a second life, with advanced speed and low noise capability allowing first pass design success in demanding RF and video applications. Today communication equipment is a large and perhaps the largest fraction of the world's semiconductor manufacturing and thus RF simulation, compact models and design kits are more common. Design tools in the silicon oriented tool flows have improved considerably, allowing more accurate non-linear simulations using periodic steady-state and harmonic balance techniques. These same techniques allow more accurate prediction of noise in non-linear and time-varying circuits, particularly mixers and oscillators. Most commercial applications today require operation up to 6 GHz, but given that a plethora of new standards are emerging with applications up to 10 GHz for UWB, 24 GHz, and 60 GHz and beyond, there are many hurdles that need to be overcome in order to create the right design methodology for silicon-based circuits.

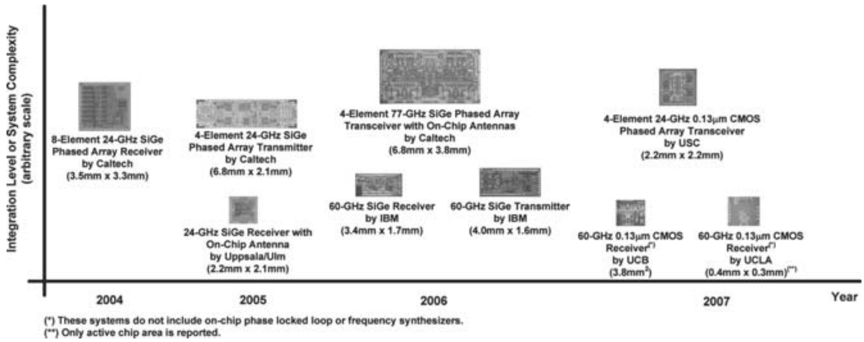

Fig. 1.4 Die photos of several published mm-wave systems integrated into silicon (from 2004-2007).

The main premise behind using silicon at millimeter waves is the higher level of integration offered at a high yield that leads into lower cost systems. Over the relatively short span of five years, several highly integrated and complex millimeter wave systems have been reported by academia and industrial research labs (Fig. 1.4)[2]. These fully integrated chips consist of several thousand RF and digital transistors and on-chip passives in multi metal-layer silicon processes and include all the receiver, transmitter and even transceiver building blocks such as low noise amplifiers, mixers, voltage controlled oscillators, phase locked loops, power amplifiers, and in some cases on-chip antennas. Moreover, in many cases multiple receive and transmit paths are integrated in a single chip to realize fully integrated phased arrays. A few trends can be predicted for the near future. The first one is the realization of complete transceivers at 60 GHz and above using a standard CMOS technology. The second direction is realizing integrated CMOS beam-forming arrays at millimeter waves. The third trend is the incorporation of more functionality into millimeter wave transceivers such as mismatch, I/Q, and array calibration, modulation and demodulation of high constellation modulation formats, and possibly wideband channel equalizers. Finally, the issue of packaging and integration of antennas with silicon chips, especially in the case of arrays, is an open area for future research. Moving into higher frequencies, 94 GHz and the near THz region, are all extremely exciting research directions that can benefit not only the engineering and scientific community, but also several industries. Overall, we should anticipate an even higher level of research enthusiasm towards silicon-based integrated mm-wave systems in the future.

[2] This figure is intended to illustrate the advancements in the silicon integration of millimeter wave systems. It is not by any means comprehensive to all the published work in the area of silicon and CMOS based integrated millimeter wave systems.

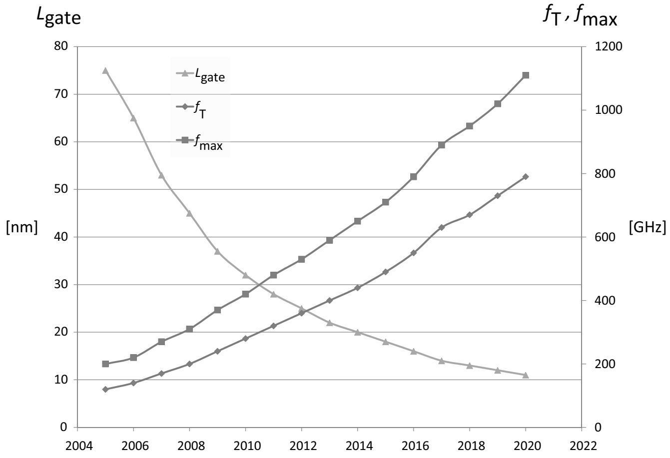

Fig. 1.5 The f_T/f_{max} trend with scaling for CMOS technology according to the ITRS 2006 [3].

1.3.1 Why CMOS?

As duly noted, CMOS technology is driven by a mass consumer market for high-speed digital microprocessors whereas specialized technologies are only needed in niche applications where the cost can be justified. For these reasons CMOS technology is a compelling choice, despite the added difficulties in doing mm-wave in a digital process.

A concerted effort by academia, industry, and worldwide research organizations has resulted in aggressive and steady technology scaling of silicon CMOS technology. This scaling was fueled by the demand for digital computation and memory and thus the technology has primarily evolved to serve these markets. The technology trend curve (f_T) shown in Fig. 1.5, shows the steady expected improvement in the operation speed of these devices. Even though specialized versions of bulk CMOS processes have been adapted to serve the communication sector for RF and microwave applications, the core devices are still digital in character. Silicon technology has been favored for scaling due to the simplicity of fabrication, which results in part from the native SiO_2 oxide and self-aligned gate formation. Scaling was initially guided by Dennard's Law [4], which called for constant field scaling, which in turn required lowering the supply voltage and also scaling the gate oxide thickness. Eventually constant field scaling was violated and high field effects, such as velocity saturation, led to additional innovations such as non-uniform doping (halo implants) and lightly doped drain contacts extensions (LDD). Despite many dire predictions about the scaling of bulk CMOS technology, the shrinking has continued down to the 45 nm technology node.

But today's CMOS is not "pure" in the traditional sense as the technology has evolved from requiring a handful of materials (Si, O, Al, and a few dopants and

rare earth metals) fabricated using optical lithographic techniques to a sophisticated process using a wide variety of materials, high-K gate dielectrics, metal gates, low-K interconnect dielectrics, seven or more metal layers incorporating Cu, and various technologies that introduce strain into the channel for enhanced mobility (SiGe drain/source or capping layers). Today devices with $L_g < 32$nm (sub optical) are routinely fabricated for high volume production. Beyond this gate length, new device structures have been proposed, such as multi-gate transistors (FinFET) and SOI transistors. In this book we limit our focus to the bulk CMOS transistor with experimental results down to the 90nm node. These devices are commercially available, relatively low cost in volume production, and perhaps represent the “sweet spot" for operating in the mm-wave band.

The technology scaling of the channel gate length L_g has resulted in raw performance benefits, particularly in the speed of the transistor. But a big penalty has been paid in other performance metrics, in particular the device output conductance g_o which translates into low intrinsic gain ($A_v = g_m/g_o \sim 10$) and extremely low supply voltages. Fortunately the supply scaling has stopped at 1V, but even at this level there are severe limitations in the achievable dynamic range of amplifiers and other key building blocks. Naturally this also translates into poor output power capability. Finally, in all technology nodes the bulk substrate is conductive which results in higher losses in passive components such as inductors and transmission lines. Fortunately, in many mixed-mode process nodes the substrate is only moderately conductive ($\rho \sim 10\Omega$-cm) to minimize substrate coupling, which is quite tolerable in many applications. Despite these shortcomings, CMOS technology has transitioned from a low performance digital process to the prevalent technology for consumer RF applications. The low gate leakage has also been a key for realizing switched capacitor analog circuits, making CMOS the technology of choice for mixed-signal circuits. In the RF regime CMOS has proved viable as a low cost alternative to SiGe and GaAs for consumer applications up to 5 GHz. Niche applications such as power amplifiers still favor technologies with higher breakdown voltages but cost considerations often dominate performance in consumer applications. For example, in many cost sensitive products, given the choice between two power amplifiers, the difference between the cost will lead to the adoption of a CMOS or silicon PA with considerably lower efficiency.

As Moore’s Law has recently pushed the f_t and f_{max} of CMOS transistors above 100 GHz, an all-CMOS solution at 60 GHz is feasible. It should be noted that the major reason behind Moore’s Law and CMOS technology scaling is a reduction of the cost per function. However, for more than forty years, CMOS technology scaling has also given us a speed improvement, leading to mm-wave CMOS designs. But the die area of a mm-wave chip is dominated by passive devices rather than active devices, so there is no inherent area advantage in using scaled technology nodes. And while Moore’s Law has enabled CMOS mm-wave circuit design, CMOS is by far not optimized for mm-wave performance. Although the success of RF-CMOS products has steered the focus of CMOS technology to consider RF design, CMOS remains in the first place a digital technology with a low supply voltage (due to breakdown), and a lossy substrate. Furthermore, operating in the 90nm node means operating

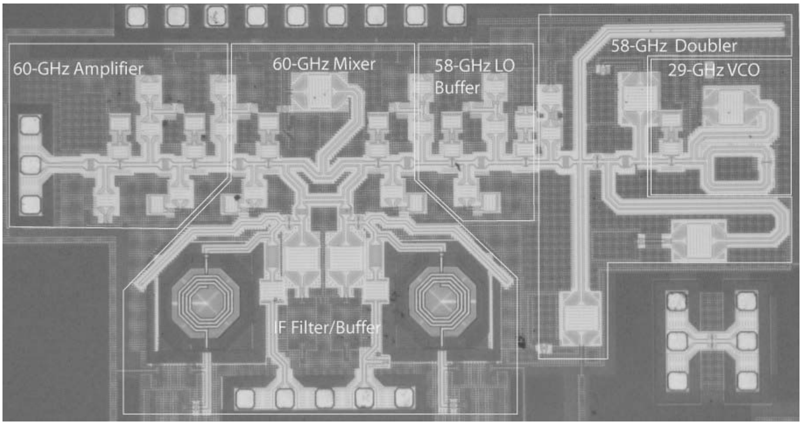

Fig. 1.6 A front-end 60 GHz receiver realized in a CMOS 130nm process technology node.

relatively close to the limits of activity ($f_T \sim 100\,\text{GHz}$, $f_{max} \sim 200\,\text{GHz}$). Therefore, circuit techniques are mandatory to achieve acceptable mm-wave performance from a CMOS technology.

While today's silicon technology is capable of operating at 60 GHz and beyond, there is still incentive to continue to use newer scaled technology since smaller transistors will provide higher performance at constant or even lower power levels. But this begs the question as to the true cost of CMOS technology, in particular if the die area continues to be dominated by the passive components.

1.3.2 True Cost of Silicon mm-Wave

It is important to briefly justify silicon technology for mm-wave applications based on economic arguments. It is now clear that the technology is capable of operating in these frequencies, but is there really a cost advantage? In particular, we have already noted that the die area of a mm-wave chip is dominated by passive devices. This is clearly visible in the die photo shown in Fig. 1.6, where the transmission lines and capacitors occupy most of the area. In this example, scaling from 130nm to 90nm would not reduce the die area noticeably. Also, the fabrication costs of silicon technology below 100nm are prohibitively high due to the mask costs, which means that silicon is only a viable solution in volume applications, where mask costs are shared over millions of parts. This also means that there should be additional investments in the modeling and tools of these devices to ensure "first pass" success in the fabrication, since any errors in the design will translate into millions of dollars of lost revenue for new masks. These problems are not new and are faced by RF and microwave circuit designers in the cellular and WiFi markets. In particular, the use of SiGe over CMOS is sometimes justified in large part because of more

reliable models and tools rather than performance benefits. Moreover, since the cost of older generation lithography tools can produce a SiGe transistor on par with a more expensive cutting-edge CMOS process, the cost difference between CMOS and SiGe is somewhat artificial. Advanced packaging makes multi-chip solutions viable and attractive. But if the entire design can be done in CMOS, then there is clearly a benefit in offering a smaller footprint product. Furthermore, the investment in modeling and designing in a cutting-edge CMOS process will pay back in the long run, since future products can leverage this knowledge and these skills as the technology cost reduces over the years.

Die costs are only a fraction of the overall cost to produce integrated circuits. Costs are often dominated by packaging and testing parts. This is particularly important for RF circuits, which require expensive test equipment and longer testing to validate functionality. Special packages with low inductance are also needed in power amplifier and ultra high frequency applications, making the die cost in some cases negligible in the overall equation. But this situation changed quickly when CMOS and SiGe designers learned to leverage automatic tuning and digital calibration techniques, particularly with the aid of advanced digital signal processing available either for “free" (shared with digital baseband) or at marginally higher costs. Today many aspects of an RF transceiver are tuned automatically, allowing lower performance technologies such as digital CMOS to compete effectively with more expensive SiGe or higher performance processes. Examples include I/Q gain/phase matching, filter passband tuning, and VCO/PLL frequency calibration. Many of these settings are done “live" to compensate for temperature and lifetime aging variations. Fully integrated transceivers are tested in loop-back mode to avoid expensive RF test equipment. With only DC probing, a transceiver chip can be tested from “bits *in* to RF *out* and then to bits *out*” from the baseband DAC to the baseband analog and RF transmitter to the RF front-end receiver blocks all the way back through the receiver analog and baseband sections. This end-to-end functionality is tested with the aid of a digital baseband.

Communication algorithms today make intensive use of signal processing to deal with multi-path fading in increasingly broader bandwidths of spectrum. The use of OFDM has become common in many existing WiFi chips (802.11a/g) and is proposed for future 4G cellular networks and short range personal area networks (OFDM-UWB adopted by Zigbee and Bluetooth). Furthermore, sophisticated Turbo codes and low-density parity codes (LDPC) are used and proposed in future systems. This means that end-to-end testing will require a very powerful DSP engine to assist in the testing, which make it more difficult to perform this level of testing in a multi-chip solution. If testing has to be performed after packaging, the yield will suffer since a faulty RF section means that the digital baseband is wasted or vice versa.

As we look at the emerging mm-wave applications, it is clear that CMOS and SiGe will only offer an advantage over traditional mm-wave technologies if packaging and testing costs can be reduced. The same approach to testing a 5 GHz part can be applied to a fully integrated 60 GHz transceiver. And while low cost packaging incurs a heavy penalty in routing signals on and off the chip, about a 1dB of loss in transmit power and 1dB higher noise figure, many low to medium performance applications will

tolerate this inefficiency in favor of lower costs. Low cost packaging has already been demonstrated for mm-wave systems [5] and many existing testing procedures for RF applications can be adopted for mm-wave. In fact, a mm-wave transceiver can be viewed as nothing but an RF transceiver with a simple front-end to translate the signals from say 60 GHz down to 1 GHz where traditional signal processing techniques prevail. Of course there are new challenges, such as the processing of extremely wideband signals (compare 2 GHz bandwidth for a 60 GHz WLAN to 200 kHz bandwidth of a cellular system). Also the incorporation of multiple antennas in mm-wave systems necessitates innovative and inexpensive ways of testing mm-wave components, especially for radar and antenna array applications. One should not doubt that these innovations will come and the ultimate cost of mm-wave silicon will reduce dramatically, enabling many new and exciting applications.

1.4 Communication in the 60 GHz Band

Many applications require or benefit from high data rates far exceeding the capability of existing WiFi and UWB technology. High quality video signals require data rates exceeding several Gb/s. This is because sending uncompressed data greatly reduces power overhead for encoding and decoding video. Set-top boxes and digital video cameras are obvious applications for this technology. In general, the need for bandwidth is insatiable, much like the demand for CPU speed, static and dynamic RAM, flash memory, and external hard disk capacity. While new spectrum is available in the low-GHz bands, these bands are likely to be overly congested in the near future. Moving up to higher frequency also provides natural isolation from fast switching digital circuitry typical of today's microprocessors, already operating at several GHz clock speed. Furthermore, the only way to extract more information from a fixed bandwidth at lower frequencies is the application of more complicated modulation schemes. To extract 1 Gb/s from 100 MHz of bandwidth obviously requires 10 bits per Hz, but only 1 bit per Hz from a 60 GHz solution with 1 GHz bandwidth. The 60 GHz system uses a relatively narrowband signal, and low order constellations can be used to transmit and receive the data. The lower GHz system, on the other hand, must use sophisticated signal modulation, often placing stringent demands on the phase noise and power amplifier linearity (particularly for OFDM), and this translates into a system with less overall sensitivity. Much energy must be consumed in the baseband of these systems to provide FFT and equalization functionality, which will end up consuming more energy per bit than a mm-wave solution, despite the higher power consumption of the front-end blocks at 60 GHz.

The key motivation for the exploration of the 60 GHz spectrum is the availability of 5–7 GHz of unlicensed bandwidth, with numerous obvious advantages and applications. This is not only true in the US but also globally. There is thus great incentive to design wireless systems capable of exploiting this bandwidth. In July 2003, the IEEE 802.15.3 working group for WPAN began investigating the use of the 7 GHz of unlicensed spectrum around 60 GHz as an alternate physical layer

(PHY) to enable very high data-rate applications such as high-speed internet access, streaming content downloads, and wireless data bus for cable replacement [1]. The targeted data-rate for these applications is greater than 2 Gb/s. Although the excessively high path loss at 60 GHz, due to oxygen absorption, precludes communications over distances greater than a few kms, short-range WPANs actually benefit from the attenuation, which provides extra spatial isolation and higher implicit security. Furthermore, due to the oxygen absorption, the FCC regulations allow for up to 40-dBm EIRP transmit power, which is significantly higher than what is available for the other WLAN/WPAN standards. The wide bandwidth and high allowable transmit power at 60 GHz enable multi-Gb/s wireless transmission over typical indoor distances ($\sim$10 m). Moving to higher frequencies also reduces the form factor of the antennas, as antenna dimensions are inversely proportional to carrier frequency. Therefore, for a fixed area, more antennas can be used, and the antenna array can increase the antenna gain and help direct the electromagnetic energy to the intended target. As previously noted, the directive antenna pattern improves the channel multipath profile by limiting the spatial extent of the transmitting and receiving antenna patterns to the dominant transmission path. Consequently, the delay spread and Rician K-factor of an indoor wireless channel can be significantly improved. However, directivity is both a blessing and a curse. Omni-directional antennas simplify the system design for point-to-multipoint systems, whereas directive antennas are best suited for point-to-point applications. A system employing antenna arrays with adaptive electronically-steerable beams will allow for mobility and ease the setup of the device compared to one using a fixed high-gain directional antenna.

1.4.1 Beam Forming

First to see that high antenna gain is a necessity in a mm-wave link, consider the simple link budget shown in Table 1.1, where we assume that we wish to communicate over a 1 GHz bandwidth given a transmit power of 10 dBm and a receiver noise figure of 10 dB, both reasonable numbers for silicon technology. The receiver noise floor is already -74 dBm for 1 GHz of bandwidth ($kTBF$), which means that we can only tolerate a path loss of about 74 dB if we assume a required SNR of 10 dB ($P_{TX}\mathcal{L} > kTBF$) . This is a very small range for an omni-directional antenna, even for line of sight propagation (about a meter). As evident in the link budget calculation, there is no link margin under the assumption of 10dB shadowing loss.

To overcome the challenge of high path loss, we must either boost the transmit power, very difficult to achieve in a low voltage process at mm-wave frequencies, or use a different antenna technology. The high path loss results from the small capture area of an antenna at 60 GHz. Assuming that the radiation flows omni-directionally, then a “ball" of radius R has an average energy density of $P_t/(4\pi R^2)$, which is simply the transmitted power divided by the surface area of the ball. We know that an antenna’s effective cross section or capture area is proportional to its area. For a simple dipole or loop antenna, this is on the order of λ^2, which means that operating

Table 1.1 Link budget for 1 Gp/s 60 GHz wireless system at 1m communication distance.

Component	Contrib.	Running Total	Comment
Tx power	+10 dBm	+10 dBm signal	PA at P_{-1dB}
Tx Ant Gain	+2 dB	+12 dBm signal power	
Path loss	-68 dB	-56 dBm signal	path loss at 1 m
Shadowing loss	-10 dB	-66 dBm signal	
Rx Ant Gain	+2 dB	-64 dBm signal	
Background noise	-174 dBm/Hz	-174 dBm/Hz noise	kT at room temp
Noise BW	+90 dB	-84 dBm noise	1 GHz noise BW
Noise figure	+10 dB	-74 dBm noise	NF of receiver
SNR at input		10 dB	Signal power / noise power
SNR required		10 dB	Coherent FSK
System Margin		0 dB	BER = 10^{-3}

at higher frequencies will incur a penalty if we use the smallest possible antennas. This result also comes using Friis' equation, the ratio of the power received P_r to the power transmitted P_t for a line-of-sight link at a distance r from the source is given by

$$\frac{P_r}{P_t} = \frac{D_1 D_2 \lambda^2}{(4\pi r)^2}$$

Here we account for the antenna directivity at the transmitter and receiver. We know for simple dipoles, the antenna directivity factor is on the order of unity and fixed for a simple antenna structure. The path loss is therefore proportional to the wavelength λ squared. This limits the range of a wireless communication system, especially when we consider the poorer noise performance of silicon at mm-wave and the limited output power capabilities. Highly directive antennas can be used to decrease the path loss, say in the form of dishes or horns, since these structures have larger capture areas. But these options are not suitable or appropriate for consumer applications. A more promising and exciting alternative is an antenna array.

A different perspective comes about if we consider that any consumer application will limit the antenna footprint to a fixed area. For a fixed antenna aperture size A, the directivity is simply

$$D = \frac{4\pi A}{\lambda^2}$$

and it can be seen that there is actually an improvement in the received power by moving to higher frequencies for a fixed antenna form factor. For example, a 60 GHz system with a 16-element antenna array occupies about the same area as a dipole antenna at 5 GHz, while providing much higher directivity.

Antenna arrays, shown schematically in Fig. 1.7, are a collection of antennas that can be used to cover a larger physical area. If variable gain and electronic phase

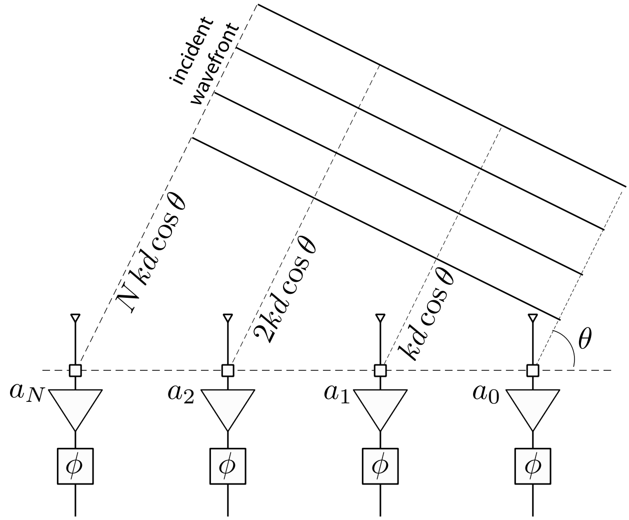

Fig. 1.7 A generic beam forming system employs arrays of antennas and transceivers to boost gain, power, and sensitivity.

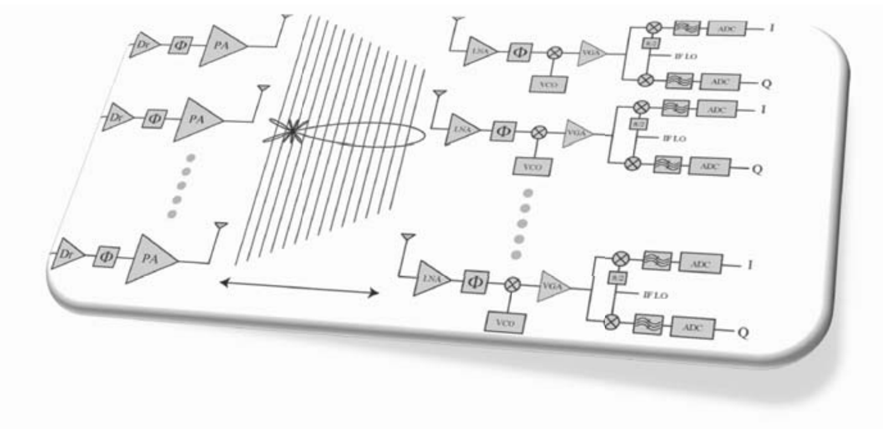

Fig. 1.8 A multiple antenna transceiver architecture employing antenna steering for maximum gain.

shifters (or time delay elements) are also incorporated into the array, then any antenna pattern can be devised, only limited by the number of antennas and a discrete Fourier transform (over the spatial dimension). A particulary important case is a “beam" pattern which is synthesized by adding enough phase shift at each element to receive an incident signal at an angle of θ coherently. Due to the small wavelength, antenna arrays can be realized in the area of the package or printed circuit board. A generic transceiver, therefore, takes the form of a multi-antenna and multi-transceiver array shown in Fig. 1.8. Beam forming improves the antenna gain while also providing spatial diversity and thus resilience to multi-path fading. The main benefit of the

Table 1.2 Link budget for phased-array 1 Gp/s 60 GHz wireless system at 1m distance.

Component	Contrib.	Running Total	Comment
Tx power	+10 dBm	+10 dBm signal	PAs at P_{-1dB}
Tx Ant Gain	+2 dB	+12 dBm signal power	
Array Gain	+9 dB	+21 dBm signal power	8-fold antenna array
Path loss	-68 dB	-47 dBm signal	path loss at 10 m
Shadowing loss	-10 dB	-57 dBm signal	
Rx Ant Gain	+2 dB	-55 dBm signal	
Array Gain	+9 dB	-46 dBm signal	8-fold antenna array
Background noise	-174 dBm/Hz	-174 dBm/Hz noise	kT at room temp
Noise BW	+90 dB	-84 dBm noise	1 GHz noise BW
Noise figure	+10 dB	-74 dBm noise	Projected NF
SNR at input		28 dB	Signal power / noise power
SNR required		10 dB	Coherent FSK
System Margin		18 dB	BER = 10^{-3}

multi-antenna architecture used here is the increased gain that the directional antenna array pattern provides. This gain is needed in order to support data rates approaching 1–10 Gb/s at typical indoor distances. Such an antenna array also allows spatial power combining which greatly simplifies the design of the transmitter. Automatic power control can be realized with no additional burden or efficiency loss by simply turning off some of the transmit paths.

To realize an antenna array, the second major challenge is providing low-cost circuit building blocks to realize the transceivers. Since multiple transceivers are potentially needed in many architectures, the power consumption of these blocks is important. Since package and board parasitics are prohibitive at 60 GHz, these transceivers should be fully integrated solutions. Given the lossy substrate is the limiting factor in low-GHz designs, it may seem that scaling to higher frequencies should be even more difficult, but in fact we shall show that relatively good passive devices are possible, even in mm-wave frequencies using inexpensive digital CMOS technology.

Table 1.2 is the link budget for a 60 GHz system communicating with 1 GHz of bandwidth. As already noted, the path loss is the dominant factor, requiring a directive antenna array to compensate for the loss of gain. Also note that the large input bandwidth increases the noise floor considerably compared to a narrowband system. Under similar conditions as the single transceiver, we see that the array allows for a substantial increase in the link margin, which allows a longer communication range or increased robustness over multi-path fading.

The third major challenge is the realization of a baseband system to filter, digitize, equalize, and detect wideband signals marred by high path loss and multi-path fading. The 802.15.3c committee has identified several channel model scenarios including

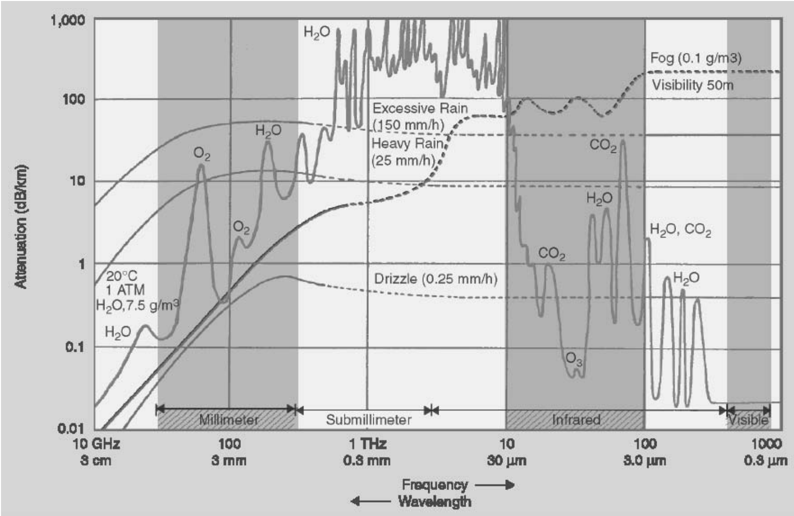

Fig. 1.9 The propagation attenuation characteristics [dB/km] versus frequency (wavelength) for Earth's atmosphere under various conditions [7] (© IEEE 2003).

indoor and outdoor, line of sight (LOS) and non-LOS (NLOS), and many variations including desktop, office, and media environments. In high data rate systems, the choice of baseband architecture can have a large impact on the overall system complexity and power consumption of the mobile transceiver. In typical "mostly digital" wireless receivers or transmitters, high resolution interface circuitry (ADCs and DACs) are required to convert the signal waveform between the analog and digital domain so that the subsequent or preceding digital processing is not limited by accuracy of these interface circuits. In multiple-antenna systems, where there may be several instantiations of the baseband circuitry, the aggregate power consumed by these high-speed, high-resolution interface circuits may become prohibitively large. It is therefore important to draw the boundary between digital and analog carefully. Similar to high speed links, analog equalization can help mitigate multi-path effects, lowering the required resolution of the ADCs. Digital estimation and control of the analog circuitry is important in realizing a robust and adaptive system [6].

1.5 Unique mm-Wave Applications

1.5.1 mm-Wave Spectrum

The availability of inexpensive small footprint silicon mm-wave transceivers leads to the possibility of higher complexity mm-wave systems incorporating dense arrays of transceivers incorporating sophisticated multi-antenna signal processing. With the

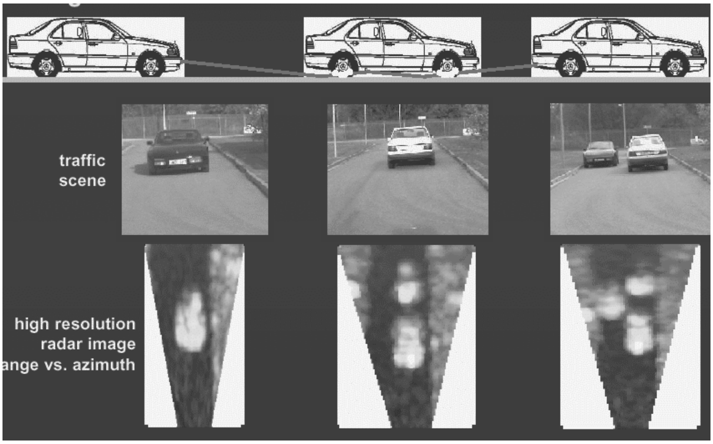

Fig. 1.10 Images produced by an automotive radar system at 77 GHz. [8] (© IEEE 2002)

continued scaling of nanoscale CMOS technology and the resulting increase in intrinsic device f_T, there are new opportunities for circuitry operating above 60 GHz. New potential applications include mm-wave imaging and sub-THz chemical detectors, with applications in astronomy, chemistry, physics, medicine, and security. Important frequencies of interest will include 90 GHz, 140 GHz, and frequencies above 300 GHz or in the so called THz regime. The reason to focus on these frequencies is evident in Fig. 1.9, which shows the attenuation of signal propagation in air for various frequencies. Clearly there are windows of opportunities where the attenuation is either minimal or maximal. The 60 GHz band is particularly lossy due to oxygen absorbtion, making it suitable for short range networks with good spatial isolation. But other bands such as 90 GHz are ideal for long range imaging. Compare the performance of 90 GHz to the visible spectrum and you can see why plants and animals evolved to use this spectrum for energy scavenging and imaging. But also compare the attenuation in the case of fog, and you can see that while a mm-wave system can outperform an optical system under (by definition) low visibility conditions[3].

1.5.2 Automotive Radar

An important application for imaging is the automotive radar operating at 24 GHz and 77 GHz. Today only luxury automobiles are equipped with mm-wave radar technology. This can aid in driving in low visibility conditions, especially in fog, and in automatic cruise control and even automated driving on a future freeway [9]. In the U.S. 42,000 people die annually as a result of automotive accidents, while

[3] That is low visibility for animals that rely on the optical spectrum for their visual needs.

1.5 million people are injured, making radar technology very attractive to realize in every automobile. Advanced video imaging, radar, GPS, and gyroscopes can be used in conjunction to give the driver a much richer set of data and ultimately a better and safer driving experience. An example image captured by a radar is shown in Fig. 1.10, where the visual image is compared to the radar image. It's clear that the radar image can look under vehicles and provide even more data than possible otherwise, especially with the abundance of larger passenger vehicles (SUVs and mini-vans) on today's roads. The key development has been the dramatic reduction in the area, cost, and power to perform the related signal processing for an automotive radar system.

Of the more than 42,000 fatalities caused by automobiles every year, about 5,000 are pedestrian fatalities. Every year, there are over 3,000,000 crashes, 1,500,000 injuries, and 9100 fatalities that occur at U.S. intersections, costing us about \$124 billion. This is one-third of all crashes and 17% of all highway fatalities. Recent accident statistics indicate that $\sim$56% of all crashes in 2005 occurred at or less than 40 mph speeds. In addition, while crashes remained flat, there was an increase in the accidents suffered by pedestrians and cyclists. Sensors that detect objects in near ranges (approximately 30 meters with large field of view) will enable many active safety applications that can significantly reduce crashes at these speeds. In Fig. 1.11 we see a host of sensors that provide object detection and parking assist, side-impact pre-crash detection, blind spot assistance, and long range object detection capability. Each sensor requires a different and custom tailored technology to meet the required specifications.

One promising technology for short range vehicular sensing applications is radar. Compared to lidar and video technologies, radar is an all-weather sensor with a sufficient resolution ($<$ 5cm) and operates in real time. In a typical radar, a radio frequency (RF) signal, usually in the form of a time-limited pulse, is transmitted towards the target(s) of interest. Information regarding the shape, distance, and speed of the target(s) is embedded in the arrival time and shape of the reflected or scattered signal(s). The radar azimuth resolution is inversely proportional to the carrier frequency. This is due to the smaller beam size of a given radiating or receiving aperture at higher frequencies. Therefore, the FCC has allocated the 76-77 GHz frequency band for long range (100m) automatic cruise control (ACC) automotive radar applications. These radars are currently realized using compound semiconductor technologies and limited to higher end cars. Radar range resolution is inversely proportional to the bandwidth of the transmitted pulse. Therefore, the FCC has allocated a wide frequency spectrum around 24 GHz (22-29 GHz) for short range automotive radar applications. Many other countries including the European Union have also approved, albeit temporarily, this frequency band for commercial vehicular use. The FCC allocated frequency band allows using the ultra wideband (UWB) technology to achieve a higher resolution for short range vehicular sensing applications such as blind spot detection, side and rear impact sensing, blind spot detection, and stop-and-go. One desirable objective is pedestrian detection and protection. The FCC defines ultra wideband (UWB) as signals having more than 500 MHz of instantaneous bandwidth or exceeding 20% fractional bandwidth.

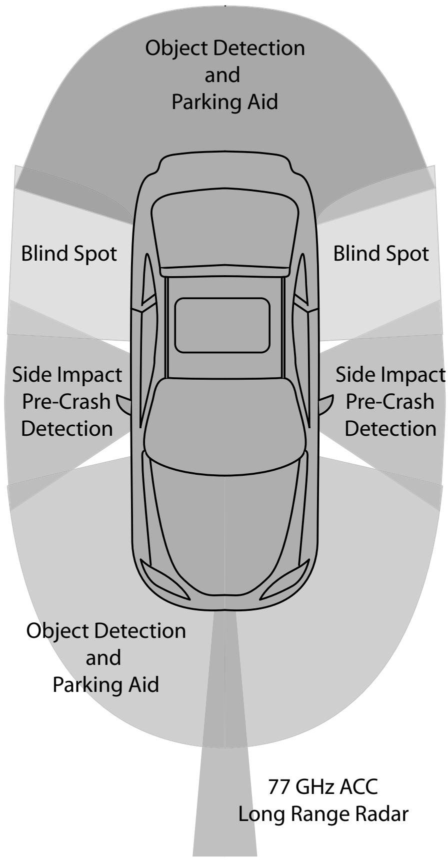

Fig. 1.11 Top view of an automobile and the desired sensors.

For instance, an UWB radar that utilizes pulses with 3 GHz of bandwidth within the FCC allocated 22-29 GHz frequency spectrum for automotive radar applications will have a range resolution of 5 cm. As the vision of the automotive industry is to incorporate several of such short-range sensors around all cars to provide 360° awareness for the driver, cost and power consumption of these sensors are crucial. Commercial SiGe UWB short-range radars at 24 GHz have already entered the market. It should be only a matter of time before CMOS versions become industry standards.

One critical system metric in all radars is the rate of successful detection and false alarms. This is even more important in automotive radars since the data collected from multiple sensors will be used to assist the driver and in some cases take the control away from the driver in critical situations. False alarms are common in typical driving situations that involve metallic bridges, ground holes, underground pipes, and parking structures. The sensor sensitivity suffers even more in a dynamic driving environment with several stationary and moving objects. The primary reason for the high rate of false alarms is the wide field of view of sensors; the radar transmitted pulse

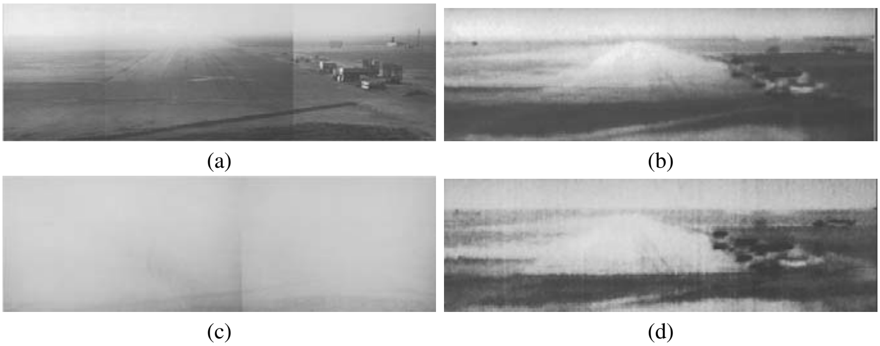

Fig. 1.12 (a) Images of an airport runway produced by an (a) optical and (b) passive mm-wave imager. Same scene under low visibility (fog) in (c) and (d) [7] (© IEEE 2003).

reaches and reflects off all the objects that are within the field of view. Therefore, the radar receives numerous pulses with varying amplitudes and shapes at different times. Advanced signal processing algorithms should be used to discriminate various targets. However, even the most advanced algorithms that are known are limited in their ability to reliably identify and track all objects in real time. Scanning radars with a limited instantaneous field of view can greatly mitigate this problem and are commonly used in existing military systems. Mechanically scanning radars are not suitable for vehicular systems due to the slow scanning rate, higher cost, difficulty in automotive sensor integration, and higher failure rate due to mechanical weariness and damages.

Phased array is a well-known technique that allow for electronic steering of the electromagnetic wave. A phased array imitates the behavior of a mechanically scanned antenna whose bearing can be adjusted electronically. A phased array transmitter can form a narrow beam that can be steered towards intended angles in a narrow field of view. A phased array receiver is only sensitive to the reflected signals that arrive at a specific angle that can be electronically controlled. Therefore, phased arrays, also known as electronically scanned arrays or beam-forming arrays, reduce the undesired effect of multiple reflections and interference while allowing for a full spatial coverage. In addition, phased array receivers enhance the radar SNR and hence its range. RF phased arrays have been extensively used in military applications for high resolution ground- and ship-based as well as airborne radars. Silicon technology, CMOS in particular, allows for the integration of an entire phased array radar with a typical size of 4 to 16-elements in a single chip. Silicon integration also improves the radar robustness and reliability through on-chip signal processing and calibration that is offered at little incremental cost. It is likely that RF and electrical engineers will play a central role in the safety of future road transportation systems.

Fig. 1.13 A passive mm-wave image can easily denude a person to reveal a hidden weapon. See optical image for a comparison [7] (© IEEE 2003).

1.5.3 mm-Wave Imaging for Medical Applications

Another potential application for mm-wave technology is passive mm-wave imaging. By detecting only the natural thermal radiation of objects in the mm-wave band, images of objects can be formed in a very similar fashion as in an optical system. Either a group of receivers or a movable mechanical antenna is required to scan the area of interest. Unless special techniques are employed, due to the relatively large wavelength, the resolution of this approach is limited to objects on the order of a mm. But this resolution is sufficient for many applications. In Fig. 1.12 we see the optical and mm-wave image seen from an airplane in good and bad visibility conditions. The mm-wave image is clearly able to penetrate through the fog and rain and provide a clear image. In security applications, passive (or active) mm-wave images of a person can be used to find hidden weapons. In Fig. 1.13 we see that a hidden gun or knife is easily visible due to the difference in emissivity of the body and the object. While metal objects are easy to detect by other means, non-metallic objects are also visible, which is a great advantage of the technology. Unlike X-ray based imaging systems, which can only be used with limited dosage with living organisms, passive mm-wave imaging does not use any additional radiation than what is naturally present. Even active imaging systems use photos with milli-eV energies compared to k-eV necessary for X-ray systems.

Other emerging applications for mm-wave technology include medical imaging for tumor detection, temperature measurements, blood flow and water/oxygen content measurements. These applications were under intense exploration in the past two decades but much of the research has discontinued due to the fact that these traditional systems were not able to compete with existing MRI or X-ray CAT scan systems. Due to the much larger wavelength, these systems have relatively poor resolution. As silicon technology allows larger arrays of transceivers to be realized in a small area at a low cost, we believe that many of these applications will re-emerge due to the immense potential for size and cost reduction. Furthermore, as we push into higher frequencies above 100 GHz, the wavelength becomes smaller and new application domains emerge. In particular, as we pass 300 GHz and enter the

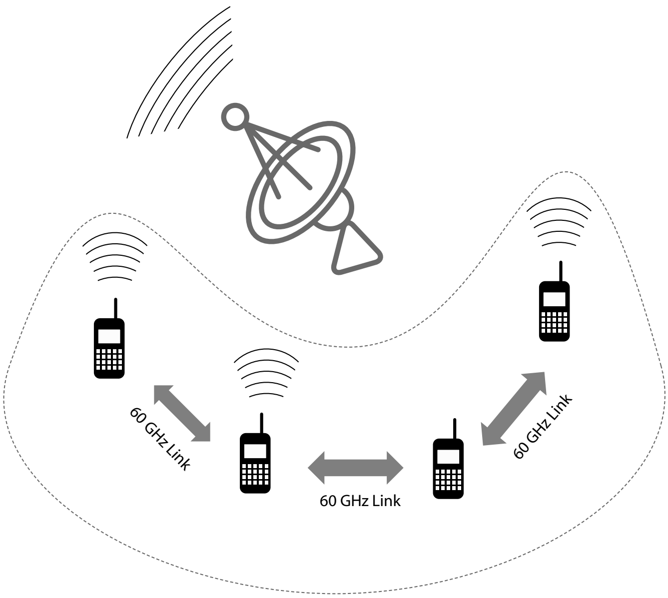

Fig. 1.14 A mm-wave collaborative distributed MIMO system.

THz regime, there are numerous applications in chemical spectroscopy due to THz vibrational resonances in molecules. One of the outstanding problems in the THz community is the lack of a sufficiently powerful transmitter. Silicon technology may offer a low power and efficient transmitter (1 mW) in the form of an oscillator and power amplifier which work directly at THz frequencies. This is in contrast to today's optical systems which use non-linear effects in crystals to produce THz emissions as a by-product.

1.5.4 Collaborative Distributed MIMO

With the wide availability of mm-wave technology, we have to fundamentally rethink the way we do communications. For instance, if a short range mm-wave link can be established with under 50 mW of power with a data rate of 2 Gb/s, then the energy consumption of such a system is about 25 pJ/bit, which compares favorably to a cellular or WiFi system consuming about 10-100 nJ/bit. Moreover, if we imagine a scenario where many portable devices with mm-wave capability are within a room or within a vicinity large enough to explore the multi-path diversity at lower frequencies, then the mm-wave links can be used to set up a collaborative and distributed MIMO system shown in Fig. 1.14. A single user forwards the data traffic to nearby users using a high bandwidth mm-wave link and then each user transmits the

same (or encoded) data to the recipient. Effectively the users together form a large aperture MIMO radio. Spatial coding is used to increase the capacity of the system greatly over a signal antenna system. This is better than a MIMO system within one device (such as 802.11n) due to the size constraints and the number of antennas one can fit into a handset footprint. Also, it is much more likely that distant radios will have higher diversity in their spatial links, allowing one to use the spectrum in an opportunistic fashion.

1.6 Overview of Book

This book develops the science and art of silicon-based mm-wave design beginning with a review of CMOS and SiGe technology considerations in Chapter 2. Next, Chapter 3 covers details of mm-wave active and passive modeling, with an emphasis on CMOS active technology. Amplifier and frequency conversion building blocks are covered in Chapter 4 in both CMOS and SiGe technology, with many examples from recent publications in conferences and journals. The design of power combining and power amplifiers is covered in Chapter 5. Voltage controlled oscillators (VCOs), frequency dividers, and other key frequency synthesizer building blocks are covered in Chapter 6. Using these foundational building blocks, the book covers system aspects of mm-wave silicon design by addressing the realization of a beam-forming array in Chapter 7.

References

1. C. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379-423, 623-656, July-October 1948.
2. S. Emami, C. H. Doan, A. M. Niknejad, and R. W. Brodersen, "Design of CMOS for 60 GHz applications," *ISSCC Digest of Technical Papers*, 2004, pp. 440 - 538.
3. http://www.itrs.net/.
4. R. Dennard, et al., "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE Journal of Solid State Circuits*, vol. SC-9, no. 5, pp. 256-268, Oct. 1974.
5. http://domino.research.ibm.com/comm/research_projects.nsf/pages/mmwave.sixtygig.html
6. D. Sobel, *Berkeley Wireless Research Center Retreat*, http://bwrc.eecs.berkeley.edu.
7. L. Yujiri, M. Shoucri, P. Moffa, "Passvie mm-Wave Imaging," *IEEE Microwave Magazine*, vol. 4, issue 3, pp. 39-50, Sept. 2003.
8. "Automotive Radar and Prospective Circuit/Antenna Technologies - From Car Collision Avoidance to Autonomous Driving," *IMS 2002 Workshop*, http://www.mtt.org/symposia/ims/2002/workshops.html.
9. http://www.path.berkeley.edu/