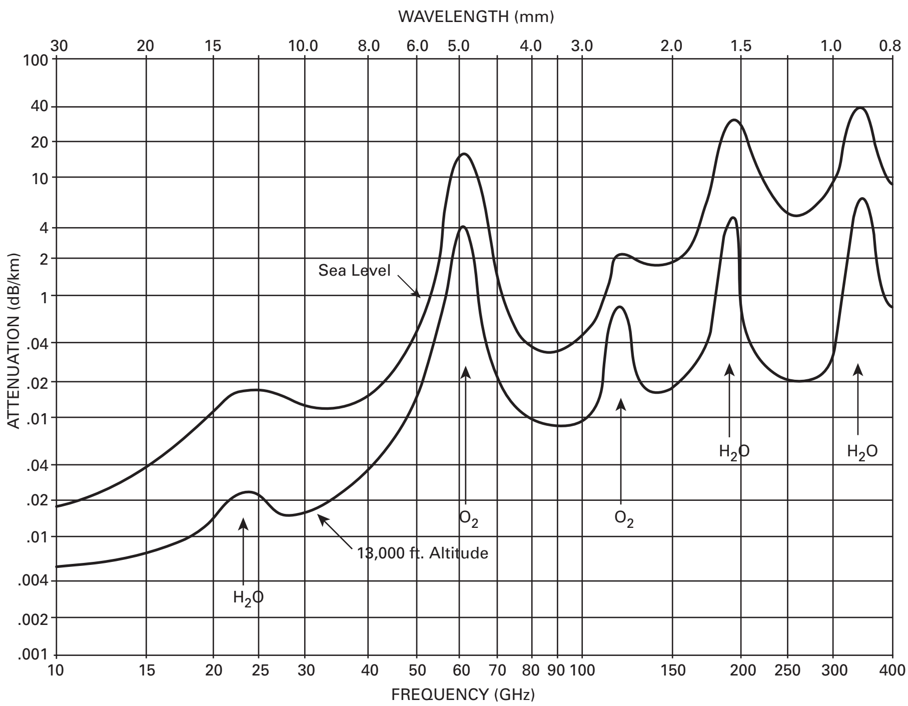
1 Introduction

Hossein Hashemi and Sanjay Raman

Advancements in semiconductor technology have led to a steady increase in the unity power gain frequency (f_{max}) of silicon transistors, both in CMOS and SiGe BiCMOS technologies. This, in turn, enables realization of complex monolithic silicon integrated circuits operating at the millimeter-wave (mm-wave) frequency range (typically defined as 30–300 GHz). Prime target applications of silicon mm-wave integrated systems include high-speed wireless access, satellite communications, high-resolution automotive radars, and imagers for security, industrial control, healthcare, and other applications. However, scaling of silicon transistors for high f_{max} comes at the expense of reduced breakdown voltages, and hence limitations on output voltage swing and power. The link range and energy consumption of wireless systems are direct functions of the transmitter output power and efficiency, respectively. Efficient generation and amplification of radio-frequency (RF) modulated waveforms using silicon transistors is an ongoing challenge due to the reduced breakdown voltage of scaled silicon transistors, loss of passive components, and the conventional linearity–efficiency trade-off. This book covers the fundamentals, technology options, circuit architectures, and practical demonstrations of mm-wave wireless transmitters realized in silicon technologies.

1.1 Why mm-waves?

The main motivation to operate the wireless systems at higher carrier frequencies is the larger available bandwidth which translates to higher data rate in communication systems and higher resolution in ranging and imaging systems. Furthermore, the size of the antenna and circuitry, typically proportional to the wavelength, reduces with increasing carrier frequency. On the other hand, operating at higher frequencies poses two fundamental challenges. First, the loss of most materials increases with the frequency; therefore, compared with radio and microwave frequencies (below 30 GHz) the electromagnetic wave at mm-wave frequencies is attenuated more as it propagates in an environment (Fig. 1.1). It should be noted that over the mm-wave spectrum there are "windows" of relatively lower attenuation around 35 GHz, 90 GHz, 140 GHz, etc., and, consequently, these bands are often selected for mm-wave applications; on the other hand, high atmospheric attenuation levels around frequencies such as 60 GHz enable more aggressive frequency reuse, and are therefore often selected for small cell or secure communications applications. Second, the performance of semiconductor



AVERAGE ATMOSPHERIC ABSORPTION OF MILLIMETER WAVES (HORIZONTAL PROPAGATION).

Fig. 1.1 Typical atmospheric attenuation in dB/km as a function of frequency.

devices worsens with frequency; this includes reduced gain, increased noise, and more nonlinearity at higher frequencies for a given technology.

Historically, mm-wave systems have been confined to defense, aerospace, and niche commercial applications due to the high cost of multi-chip-module (MCM) approaches and the need to use compound semiconductor processes to achieve required performance. However, over the past decade, there has been an explosion of research and development towards system-on-a-chip (SOC) realizations of complex wireless systems operating at mm-waves for high-volume commercial applications. Commercial complex mm-wave SOCs, such as 60 GHz phased array transceivers for high-speed wireless access, exist today [1–3]. Even certain defense applications, such as large-scale mm-wave phased arrays for helicopter operations, promise to be benefited by silicon implementations [4]. The main commercial applications being pursued at mm-waves include high-speed wireless connectivity with primary focus in the 60 GHz industrial, scientific, and medical (ISM) frequency band; high-resolution automotive radars with primary focus at the 77 GHz frequency band; mm-wave backhaul in the Ka and E bands; and active and passive RF imagers with primary focus at frequencies above 100 GHz. The fifth-generation commercial wireless standard (5G), targeting systems beyond 2020, is expected to include mm-wave operation for high-data-rate wireless access between small cells and mobile devices. High-resolution radar continues to be

IEEE Frequency Bands [GHz]	Frequency [GHz]	Applications	Applicable Standards
Ka (27-40)	28, 38	Mobile communications for 5G Cellular networks (emerging)	None yet available
	27.5-30	SATCOM uplinks (e.g., Inmarsat Global Xpress: 27.5-30) and downlinks (e.g., Iridium: 29.1-29.3)	MIL-STD-188-164, ITU-R S.524-9, FCC 25.138, ETSI EN 303 978
	24.25-30	L-3 ProVision imaging scanners at airports	IEEE C95.1
	35	Munitions and missiles seekers and sensors	Unknown
V (40-75)	43.5-45.5	U.S. AEHF military SATCOM system uplinks*	MIL-STD-3015
	57-66	"Last inch", short range wireless communications	IEEE 802.11ad, IEEE 802.11aj, IEEE 802.15c
	71-75	"Last mile", point-to-point backhaul wireless communications	ETSI TS 102 524
W (75-110)	75-76, 81-86, 92-95	"Last mile", point-to-point backhaul wireless communications	ETSI TS 102 524
	76-77	Autonomous cruise control (ACC) "long range" automotive radar	ETSI EN 301 091 parts 1 & 2
	77-81	Short range "stop & go" automotive radar	ETSI TR 102 263
	94	Missile seekers, collision avoidance radars	Unknown
	85-110	Imaging for medicine, biology, and security	IEEE C95.1
G (110-300)	110-120	Imaging for medicine, biology, and security	IEEE C95.1
	220-240	Long range wireless communications, atmospheric research radar	None yet available
	120, 183, 325	Short range wireless communications (emerging)	None yet available

* U.S. AEHF system downlink frequencies are located at 20.2 GHz – 21.2 GHz (IEEE K band).

Fig. 1.2 Summary of major mm-wave applications and applicable standards.

the primary application for mm-wave military systems. Figure 1.2 summarizes major mm-wave applications and their applicable standards.

1.2 Why silicon?

The advancement of silicon technologies, CMOS in particular, is motivated by performance gains of digital computation and signal-processing integrated circuits. Specifically, the computation speed and power consumption of digital circuits improve with technology scaling. The large investment required for advancing the silicon manufacturing technologies has been justified by the large demand due to the economy of scale. Thanks to groundbreaking research since the 1990s, today most of the RF functions of a wireless system are also realized in the same digital CMOS process leading to SOC realizations. Compared with the traditional multi-chip-module (MCM) approaches, SOCs reduce the cost, complexity, and power consumption, while enhancing robustness thanks to on-chip calibration, built-in self-test (BIST), and self-healing schemes. Furthermore, availability of "free" digital functions has enabled new system architectures with improved performance over conventional schemes.

The widespread usage of silicon technologies for complex mm-wave integrated systems is a result of large-scale research and development over the past two decades. Silicon technologies capable of operating at mm-waves were available in the 1990s [5, 6], followed by monolithic mm-wave circuit realizations shortly after [7]. Early efforts towards realization of complex mm-wave integrated circuits in standard silicon technologies were led by Caltech [8–11], IBM T. J. Watson Research Center [12, 13], UC Berkeley [14, 15], UCLA [16, 17], and the University of Toronto [18, 19] among others around the mid 2000s. Later, in addition to the aforementioned groups, several more research groups such as Georgia Tech [20], UCSD [21–24], National Taiwan University [25, 26], Intel [27], and Tokyo Institute of Technology [28] among others made significant contributions in the research and development of mm-wave silicon complex integrated systems.

While technology scaling provides transistors with higher transistor unity power-gain frequency, it also reduces the breakdown voltage and hence the maximum allowable voltage swing. In fact, there is an inverse relationship between maximum speed and breakdown voltage for a given semiconductor material. The lower allowable voltage swing degrades the signal-to-noise ratio (SNR) and linearity of many circuit building blocks, and also challenges efficient generation of high-power signals. Silicon technology does not lead to higher-performance circuit building blocks with a fixed topology when compared with a compound semiconductor realization. The main advantage of using a silicon technology for high-performance mm-wave systems, in addition to lower cost and footprint, is the higher performance of the entire system enabled by new integrated architectures that leverage combination of analog, mixed-signal, and digital designs.

Chapters 2 and 3 discuss the current state of the art in SiGe and CMOS technologies, respectively, for mm-wave transmitter applications.

Efficient, watt-level radio-frequency (typically <6 GHz) power amplifiers and transmitters now exist commercially. The choice of using silicon versus compound semiconductor technologies in an RF power amplifier depends on the specific application, market demand, and related economics. It is quite conceivable that the growth of wireless devices and connectivity at radio frequencies thanks to CMOS realizations will be repeated at mm-wave frequencies for a complementary set of applications.

1.3 Wireless communication basics

The general form of a modulated waveform, which can be the electric or magnetic field of a propagating electromagnetic wave, can be expressed in the so-called polar form as

$$x_{polar}(t) = a(t)\cos(\omega_c t + \phi(t)), \tag{1.1}$$

where ω_c is the carrier frequency, and $a(t)$ and $\phi(t)$ are the amplitude modulation (AM) and phase modulation (PM) portions of the waveform and contain the information. This expression can also be written in another form, commonly referred to as the Cartesian form, as

$$x_{Cartesian}(t) = I(t)\cos(\omega_c t) + Q(t)\sin(\omega t), \tag{1.2}$$

where $I(t)$ and $Q(t)$ contain the information and are referred to as the in-phase and quadrature-phase components, respectively. While the aforementioned two forms are mathematically interchangeable, they inspire different architectures for their generation and detection, in turn offering various implementation trade-offs. One notable difference between the aforementioned forms is that $a(t)$ and $\phi(t)$ occupy significantly larger bandwidth in the frequency domain as compared to $I(t)$ and $Q(t)$.

The main function of a wireless transmitter is the energy-efficient generation of a high-power modulated waveform as close to the aforementioned ideal forms as possible. Nonlinearity of the transmitter creates distortion in the generated waveform; moreover, noise can be added to amplitude and/or phase components as well. Distortion in the transmitted signal degrades the receiver detection, or, more specifically, the bit error rate (BER) in digital communication systems. Error vector magnitude (EVM), a metric that quantifies the deviation of constellation points in a digitally modulated signal from their ideal state, degrades with transmitter nonlinearity. Nonlinearity also causes the power of the bandlimited modulated signal to get spread to a larger bandwidth in a process commonly referred to as spectral regrowth. Spectral regrowth reduces the SNR of other users of the wireless network, and degrades the overall system capacity. Adjacent channel power ratio (ACPR) characterizes the spectral regrowth in wireless transmitters due to nonlinearity.

The maximum achievable data rate of a communication system, referred to as the channel capacity, is given by Shannon's capacity limit equation, under the assumption of additive white Gaussian noise (AWGN), as

$$C = BW \log(1 + SNR), \tag{1.3}$$

where BW is the system bandwidth. The linear increase of capacity with bandwidth is the main motivation behind moving to mm-wave carrier frequencies where more bandwidth is available. On the other hand, increasing the wireless capacity through SNR requires an exponentially larger relative signal power, which often leads to an unrealistically large transmitter power requirement. In a line-of-sight (LOS) wireless link with distance d between the transmitter and receiver, the power of a received signal, P_r, is given by

$$P_r = P_t G_t G_r \left(\frac{\lambda}{4\pi d}\right)^2, \tag{1.4}$$

where P_t is the power of the transmitted signal, G_t and G_r are the transmitter and receiver antenna gains, and λ is the wavelength. Therefore, for a system with equal transmit and receive antenna gains, the power of received signals with shorter wavelengths (higher frequencies) decreases. In order to maintain a sufficiently high received signal energy at mm-waves, the antenna gain is typically increased. This implies a more directional wireless link. Fortunately, the direction of the narrow electromagnetic beam can be steered electronically using phased-array beam-forming networks consisting of multiple antennas and the requisite control electronics. Fortunately, the smaller wavelength at mm-wave frequencies leads to a small footprint for the multi-antenna beam-forming scheme. The requirement for additional electronic circuitry in multi-antenna mm-wave transceivers necessitates a silicon SOC approach to reduce the cost, power consumption, and footprint. A major advantage of mm-wave beam-forming is that co-channel interference can be mitigated, leading to an increase in overall network capacity.

1.4 Wireless transmitter architectures

The modulated waveforms of Eqns (1.1) and (1.2) can be generated directly in the so-called direct I/Q (homodyne) and polar transmitter architectures. Direct I/Q modulators require accurate I/Q generation circuitry and linear amplification; as such, they often utilize additional feedback circuitry for I/Q correction and/or amplifier linearization. Furthermore, linear amplifiers are fundamentally inefficient. Polar modulators, on the other hand, can operate with switching amplifiers with a theoretical efficiency of 100%. However, they require accurate synchronization between the amplitude and phase signal paths that have a larger bandwidth compared with their I/Q counterparts (Fig. 1.3).

Modulation and frequency up-conversion can occur in multiple stages such as in the heterodyne architecture. Waveforms that do not carry amplitude-modulation information, also known as constant-envelope waveforms, can be created within phase-locked loops (PLL) or other forms of phase/frequency modulators, and consequently amplified with efficient switching class (nonlinear) power amplifiers (PA). Architectural trade-offs between various transmitters are well documented and will not be elaborated on here.

The increase of transistor speed has also led to digital generation and processing of RF waveforms. Modern RF transmitters, and transceivers in more general terms, use more digital building blocks for several reasons. The performance of digital circuits improves with technology scaling. Digital circuits consume less chip footprint compared with their analog counterparts. Multi-function and reconfigurable systems are easier to realize using digital circuits than with analog circuits. Digital circuits are more robust to device mismatches, as well as process, voltage, and temperature (PVT) variations. Finally, in principle, the design and verification of digital circuitry can be more straightforwardly automated compared with analog circuits. Assuming the continued advancement of transistor performance over time, it is conceivable that future high-performance mm-wave transmitters will also include significant digital circuitry at their core. Several of the chapters of this book cover digitally assisted or inspired power amplifier and transmitter concepts and circuits.

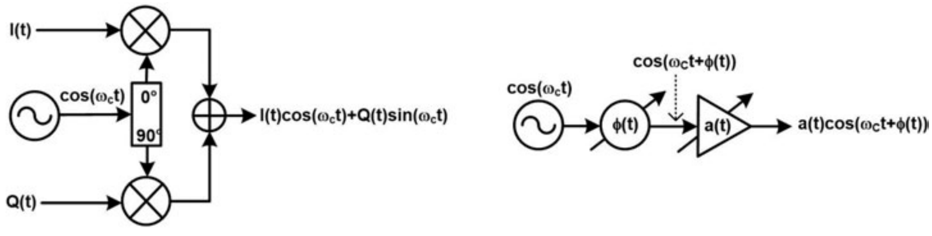


Fig. 1.3 Direct generation of modulated waveforms using direct I/Q and polar transmitter architectures.

1.5 Power amplifier basics

Historically, power amplifiers have been classified as either linear or switching amplifiers. This designation is somewhat misleading as several amplifier topologies that fall under the linear category are not strictly linear. In fact, with the exception of small-signal amplifiers (Class A topology with small input and output waveforms), all other amplifiers are nonlinear to some extent. The linearity versus efficiency trade-off of conventional (analog) power amplifiers is well documented. For instance, while the theoretical efficiency of Class A amplifiers is limited to 50%, switching amplifiers with Class D, E, F, . . . topologies can approach theoretical 100% power efficiency. Power added efficiency (PAE) measures the true efficiency of a power amplifier as it includes the AC power delivered at the input, P_{in}, as well as the power delivered through the DC supply, P_{DC}, as

$$PAE = \frac{P_{out} - P_{in}}{P_{DC}}. \tag{1.5}$$

At frequencies where the transistor can have significant power gain, the input RF power can sometimes be neglected – hence, instead of PAE, drain or collector efficiency, measuring the ratio of output power delivered to the load to the DC power drawn from the supply can be used. In fact, classical design criteria, optimization steps, and performance limits of power amplifiers do not consider the finite power gain of transistors. This assumption may not hold true at mm-wave frequencies where silicon transistors do not necessarily offer significant power gain. Several chapters in this book cover the effect of finite transistor power gain in the formulation and design steps of power amplifiers at mm-wave frequencies.

Power amplifiers can also be classified according to the main power transistor's operating mode. The output terminal of a transistor can be modeled as either a current source or a voltage switch. In single-ended amplifiers, the former applies to the forward active region of BJTs (bipolar junction transistors) and saturation region of FETs (field-effect transistors), while the latter applies to the saturation region of BJTs and triode region of FETs. Passive networks connected between the transistor terminals, and specifically at the output node, shape or scale the voltage and current waveforms. Waveform shaping is used primarily to reduce the overlap between the power transistor's voltage and current waveforms and hence increase the power efficiency. Waveform scaling is used to deliver the desired amount of power to the load given the limited breakdown limits of a transistor. Impedance transformation and power combining are the primary approaches used to scale the voltage or current waveforms.

Advancements in digital-signal-processing techniques, combined with the low overhead of complex digital circuitry, have led to widespread utilization of linearization techniques in power amplifiers and transmitters. Today, it is generally believed that, for the same specification, a transmitter can be designed to be more power efficient if linearization circuitry is combined with more efficient, nonlinear power amplifier cores, as opposed to relying on inefficient linear power amplifier cores. In fact, many modern transmitter architectures designed to support digital modulation formats now rely on switching power amplifiers. In envelope tracking approaches, the DC power delivered to the power amplifier output stage is modulated by the amplitude envelope of the waveform to be transmitted, thereby avoiding wasting DC power headroom while the signal amplitude is at lower swings.

It should be noted also that devices used in a power amplifier undergo large voltage and current swings, and oftentimes at elevated temperatures. This creates concerns over their short-term and long-term reliability. The concern is elevated at mm-wave frequencies as the transistors (with limited power gain) are typically pushed hard to deliver the desired amount of power. Fortunately, availability of “free” digital circuitry combined with advanced BIST and self-healing techniques mitigates some of the aforementioned risks. This will be elaborated on in a dedicated chapter. Power-combining strategies can also be utilized to increase the overall output power of the transmitter while maintaining more reasonable power dissipation levels for the individual transistor stages.

1.6 Examples of commercial mm-wave applications

1.6.1 60 GHz wireless communications: the IEEE 802.11ad standard

The frequency band around 60 GHz has been designated as an industrial, scientific and medical (ISM) band for unlicensed wireless communications. This band coincides with one of the resonance modes of the oxygen molecule, resulting in higher propagation loss. This higher propagation loss, while unfavorable for the usage of 60 GHz in long-range applications, enables tighter frequency reuse in a dense wireless network planning, resulting in higher aggregate capacity. The Wireless Gigabit Alliance (WiGig) was a trade association that developed and promoted the standardization of 60 GHz unlicensed wireless communications for creating interconnected home entertainment and office devices such as computers, displays, and mobile phones. The developed standard is called IEEE 802.11ad, and the WiGig alliance was subsumed by the Wi-Fi alliance. The physical layer (PHY) supports data rates up to 4.62 Gbps and 6.75 Gbps for single-carrier (SC) and orthogonal frequency division multiplexing (OFDM) modes, respectively, over 2.16 GHz channel bandwidths. The media access control (MAC) layer is compatible with the IEEE 802.11 standard.

Several research groups and industrial entities have produced chips or chipsets that are compatible with the IEEE 802.11ad standard. A few representative examples are here to provide a perspective of the technology evolution and state-of-the-art performance numbers. In 2011, SiBeam (later acquired by Silicon Image) reported 65 CMOS chips with up to 32-element beam-forming, capable of supporting the early drafts of 802.11ad standard with 3.8 Gbps data rate over 50 m (32TX/32RX) [1]. In 2013, Panasonic reported an 802.11ad CMOS chipset (90 nm RF and 40 nm BB), capable of supporting SC mode, achieving 1.5 Gb/s data rate over 1 m [2]. In 2014, Broadcom Inc. reported an 802.11ad 40 nm LP CMOS chipset with 16TX/16RX beam-forming, capable of supporting SC and OFDM modes, and achieving 4.6 Gbps data-rate (PHY rate) over 10 m [3].

1.6.2 77 GHz automotive radars

Various forms of sensors are used to enhance the safety and experience of driving. Millimeter-wave radars have been used in high-end cars for over a decade. The 76–77 GHz band has been the primary standardized frequency band for automotive radars throughout the world [29]. Current automotive long-range radars (LRRs) based on frequency-modulated continuous wave (FMCW) schemes are able to detect objects as far as 250 m away with fixed or steerable beams. There have been ongoing efforts towards utilizing more bandwidth within 77–81 GHz for automotive short-range radars (SRRs) with better range resolution for applications such as park assist.

While early realizations of automotive radars were based on discrete components, modern versions are highly integrated. A few representative examples are provided here. The advantage of silicon technologies towards realization of low-cost automotive radars has been recognized since the early 2000s [30, 31]. In 2008, a collaborative team led by Infineon Technologies reported a monolithic four-channel 77 GHz FMCW automotive radar chip capable of operating across $-40\,^{\circ}\mathrm{C}$ to $+125\,^{\circ}\mathrm{C}$ in a 200 GHz SiGe:C production technology [32]. In 2012, Freescale Semiconductor reported a transceiver chipset consisting of a four-channel receiver and a single-channel transmitter 200 GHz SiGe BiCMOS technology for 76–77 GHz LRR and 77–81 GHz SRR automotive applications [33]. More recently, UC San Diego and Toyota Research Institute have led the development of automotive phased array radars based on custom SiGe HBT chips [34, 35]. The increased interest and likelihood for commercial autonomous vehicles (self-driving cars) will undoubtedly result in further research and development towards low-cost high-performance mm-wave radars.

1.6.3 Future applications

The fifth-generation wireless communication standard (5G) is expected to include millimeter-wave for high-speed wireless access in mobile systems [36, 37]. Millimeter-waves continue to be among the legitimate candidates for high-speed connectivity in bulk-haul communications [38, 39], data centers, etc. Higher mm-wave frequencies (100–300 GHz) may be used for high-resolution imaging and sensing applications. In all of the aforementioned applications, economical as well as technical considerations will determine the advantages of a silicon solution.

1.7 Examples of military mm-wave applications and initiatives

1.7.1 Advanced EHF SATCOM

The Advanced Extremely High Frequency (AEHF) system is a US military satellite communications system developed to provide secure, jam-resistant global communications with data rates, overall throughput, and user coverage far exceeding those of the existing, legacy Milstar system [40, 41]. The planned system consists of four geosynchronous Earth orbit (GEO) satellites with inter-satellite links for global communications. The satellite downlink and uplink frequencies are 20.2–21.2 GHz and

43.5–45.5 GHz, respectively [42]. Data rates up to 8.2 Mb/s to future Army AEHF ground terminals are planned (the system will be able to support ∼6000 terminals). Since the gain of an antenna is proportional to the antenna's physical area and inversely proportional to the square of the signal wavelength, high-gain antennas can be realized in relatively small physical sizes at 45 GHz. Currently low-cost terminal (LCT) technologies are under development by various companies for SATCOM-on-the-move (SOTM) and SATCOM-at-the-halt applications. It is thus envisioned that chip-scale, watt-level, highly linear, efficient silicon transmitters may be used in compact satellite terminals.

1.7.2 The DARPA ELASTx program

In 2009, the US Defense Advanced Research Projects Agency (DARPA) recognized that watt-level transmitter output power can be achieved in silicon technologies with efficiencies significantly beyond the state-of-the-art and with linearity sufficient to support high-order digitally modulated waveforms (e.g., 64QAM).

This led to the launching of the Efficient Linearized All-Silicon Transmitter ICs (ELASTx) program [43]. Figure 1.4 shows a conceptual block diagram of an envisioned efficient linearized all-silicon transmitter IC, based on extensive leveraging of integration available with today's silicon technologies. In the envisioned SOC, transmitter inputs are baseband I/Q symbol samples generated in a digital processor. The samples are converted to analog I/Q signals and passed to the I/Q modulator, where a modulated complex waveform is created and up-converted to the carrier frequency of interest. The output of the transmitter is linear RF power, where PA linearization can be accomplished by using a variety of techniques (e.g., feedforward, Cartesian feedback, pre-distortion, polar modulation, outphasing, etc.). Watt-level output powers at mm-wave frequencies can be accomplished through power-combining strategies such as planar parallel, spatial, and waveguide power combiners, as well as series device stacking to enable overall high output swings in low-voltage scaled silicon technologies. At the same time efficiencies >50% were targeted by the program, through various

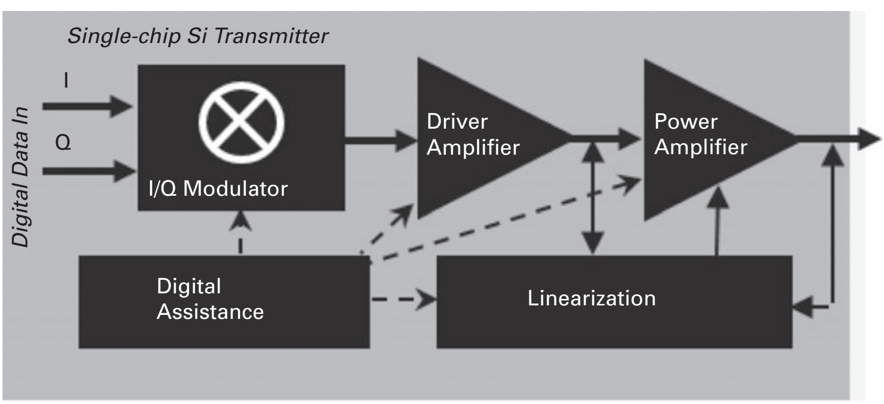


Fig. 1.4 Conceptual block diagram of an ELASTx silicon mm-wave transmitter SOC.