Abstract—In this paper, we provide a survey on techniques to improve the spectrum and energy efficiency of wireless communication systems. Recognizing the fact that power amplifier (PA) is one of the most critical components in wireless communication systems and consumes a significant fraction of the total energy, we take a bottom-up approach to focus on PA-centric designs. In the first part of the survey, we introduce the fundamental properties of the PA, such as linearity and efficiency. Next, we quantify the detrimental effects of the signal non-linearity and power inefficiency of the PA on the spectrum efficiency (SE) and energy efficiency (EE) of wireless communications. In the last part, we survey known mitigation techniques from three perspectives: PA design, signal design and network design. We believe that this broad understanding will help motivate holistic design approaches to mitigate the non-ideal effects in real-life PA devices, and accelerate cross-domain research to further enhance the available techniques.

Index Terms—Energy efficiency, power amplifier, green communications, green wireless communications, green ICT.

I. INTRODUCTION

Recently, information communications technology has been studied widely to achieve low energy consumption or high energy efficiency (EE) while meeting the high quality-of-service (QoS) and spectral efficiency (SE) requirements [1]–[5]. Compared to fixed line networks in which the communicating nodes are connected through physical wires, significantly more energy is consumed in wireless access networks, particularly at the transmitter [3], [4]. This is because the transmission energy has to be increased to amplify the transmitted signal, so as to overcome path loss and to provide a sufficient margin to random fading and interference in the wireless medium. The amplification is performed via a power amplifier (PA). The PA represents one of the most energy consuming components in a wireless system. In cellular networks, for example, energy is consumed mostly at a base station (BS) [3], of which 50%–80% of power is consumed at the PAs [6]–[8]. This high power consumption of PA is mainly due to two reasons: the limited achievable efficiency, and the limited dynamic range within which the PA can produce a linear amplification. As a result, the achievable EE and SE of the system is far from the ideal case but will depend greatly on the PA implementation [9]. Therefore, a holistic system design can effectively reduce the energy consumption incurred by the PA and improve the system performance, as evidenced by the literature that we will present.

To reduce the PA’s energy consumption, a good understanding on the PA technology is essential. Moreover, an overview and categorization of the state of the art is needed. This PA-centric survey article aims to serve both of these purposes. We believe this survey will serve as a good foundation for the wireless communication engineers who wish to develop novel and effective techniques on energy and spectrum efficient wireless communication systems. This survey complements existing surveys in green communications with broader scope that are not specifically oriented to the effects of PA, e.g., surveyed in [1], [3], [4], [10]–[13]. The summary of the survey is as follows:

Section II: In the first part of the paper, we review the fundamental properties of PA. We focus only on PA that is used for radio frequency (RF) communications. In particular, we highlight two fundamental characteristics of a real-life PA, namely, non-linearity in signal amplification and inefficiency in energy consumption. When the PA input signal is higher than the linear region threshold, the PA exhibits a non-linear property and the output signal will be distorted; on the other hand, when the PA input signal decreases from a saturation point where the maximum output power is achieved, the PA efficiency drops significantly. Readers who are familiar with the PA may skip Section II that serves as an introduction of PA.

Section III: To quantify the practical PA effects on communications, we formalize the SE and EE tradeoff, moving from the case of an ideal PA to the case of a real-life PA. The nonlinear amplification results in loss in signal fidelity as distortion is introduced to the signal, while the inefficiency results in waste of energy; hence it is most spectrum and energy efficient to operate the PA at its saturation point [85]. However, due to the dynamic range of the PA’s input signals, the PA cannot always operate at the saturation point, hence there will be a tradeoff between the EE and SE performances. This tradeoff is even more important for widely-fluctuating signals, such as multicarrier signals in orthogonal frequency division multiplexing (OFDM) or orthogonal frequency division multiple access (OFDMA) system. Since most modern wireless communication systems have high dynamic variation over time, a careful evaluation of the tradeoff is necessary.
From the discussion regarding theoretical and practical SE-EE tradeoffs, we draw insights on how the non-linearity and inefficiency of the PA separately affects the SE-EE tradeoff. This study also provides intuition and reveals the motivation on some of the current techniques that are used to mitigate the detrimental effects of PA.

Sections IV, V, VI: In the last part of the paper, which is summarized in Table I, we survey the existing technologies to resolve some of the issues that arise from the real-life PA, so as to improve the SE-EE tradeoff. This article surveys the literature over the period 1995–2013 on the PA-centric energy efficient technologies for wireless communications. The survey categorizes the energy efficient approaches into three categories: PA design, signal design, and network design. The categorization is due to the fact that the designs are typically performed by different domain experts, namely RF engineers in the analogue domain for PA design, communication engineers and signal processing engineers in the digital domain for signal design, and network and cellular engineers in the network domain for network design. The three domains display a hierarchical relationship, in the sense that the PA design has effects on the signal and network design and performance, while the signal design has effects on the network design and performance.

- In Section IV, we survey the PA design approaches including technologies that directly improve PA’s reliability, linearity, and/or efficiency, via PA architecture.
- In Section V, we review signal design approaches that exploit the knowledge of PA’s input/output signal properties and design the signals for given PA architecture. Two typical methods, namely, peak-to-average power ratio (PAPR) reduction and linearization methods, are surveyed to illustrate how the signal design implicitly affects linearity and efficiency of the PA. The linearity improvement by signal design can also improve the PA reliability as it reduces the PA saturation probability. Furthermore, the signal design may allow high input power with high efficiency of the PA.
- In Section VI, we survey two network design approaches to increase the network EE, namely, network densification and network protocol. Herein, the knowledge of network traffic and load is used to activate, deactivate, or select transmitters or PAs. Since a transmitter can be switched off with the load shifted to other transmitters or to a later time; thus, the energy wastage due to the switching on and operating of the PA at the low input signal level can be avoided. Equivalently, we can interpret the network EE improvement as PA efficiency improvement.

As summarized in Table I, each individual method has its own merits, yet there are also challenges to the implementation. For example, PAPR reduction and linearization methods in signal design incur latency due to the complex signal processing and may restrict their uses, and network densification requires additional cost for the infrastructure. We discuss the remaining challenges and future work for energy efficiency issue in Section VII.

### Table I

**PA-centric Energy Efficient Technologies: PA Linearity (L) and Efficiency (E).**

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Methods</th>
<th>Improvement</th>
<th>Challenges</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA Design (Sec. IV)</td>
<td>linear architecture</td>
<td>L</td>
<td>high cost,</td>
<td>[14], [15]</td>
</tr>
<tr>
<td></td>
<td>parallel architecture</td>
<td>E</td>
<td>large form factor</td>
<td>[16], [17]</td>
</tr>
<tr>
<td></td>
<td>switching architecture</td>
<td>E</td>
<td></td>
<td>[18]–[26]</td>
</tr>
<tr>
<td></td>
<td>envelope tracking (ET) architecture</td>
<td>E</td>
<td></td>
<td>[27]–[31]</td>
</tr>
<tr>
<td></td>
<td>envelope elimination &amp; restoration (EER/Kahn)</td>
<td>E</td>
<td></td>
<td>[32]</td>
</tr>
<tr>
<td></td>
<td>outphasing technique (LINK)</td>
<td>E</td>
<td></td>
<td>[33]</td>
</tr>
<tr>
<td></td>
<td>Doherty technique</td>
<td>E</td>
<td></td>
<td>[34]–[39]</td>
</tr>
<tr>
<td>PAP reduction</td>
<td>clipping coding</td>
<td>L, E</td>
<td>out-of-band emission</td>
<td>[40]–[46]</td>
</tr>
<tr>
<td></td>
<td>partial transmit sequence (PTS)</td>
<td></td>
<td>additional resource,</td>
<td>[47]–[50]</td>
</tr>
<tr>
<td></td>
<td>selective mapping (SLM)</td>
<td></td>
<td>latency,</td>
<td>[51]</td>
</tr>
<tr>
<td></td>
<td>tone reservation (TR)</td>
<td></td>
<td>complexity</td>
<td>[52], [53]</td>
</tr>
<tr>
<td></td>
<td>tone insertion (TI)</td>
<td></td>
<td></td>
<td>[54]</td>
</tr>
<tr>
<td>Linearization</td>
<td>feed forward</td>
<td>L, E</td>
<td>sensitive to PA,</td>
<td>[55], [56]</td>
</tr>
<tr>
<td></td>
<td>feedback</td>
<td></td>
<td>additional circuit,</td>
<td>[57]–[61]</td>
</tr>
<tr>
<td></td>
<td>digital predistortion (DPD)</td>
<td></td>
<td>complexity</td>
<td>[62]–[65]</td>
</tr>
<tr>
<td>Network densification</td>
<td>in/out band small cells</td>
<td>L, E</td>
<td>infrastructure,</td>
<td>[66]–[71]</td>
</tr>
<tr>
<td>(Sec. VI)</td>
<td>distributed antenna system (DAS)</td>
<td></td>
<td>overhead signalling,</td>
<td>[72]–[75]</td>
</tr>
<tr>
<td></td>
<td>cooperative communications (relay)</td>
<td></td>
<td>scalability,</td>
<td>[5], [76], [77]</td>
</tr>
<tr>
<td></td>
<td>Network protocol</td>
<td>E</td>
<td>handover,</td>
<td>[78]</td>
</tr>
<tr>
<td></td>
<td>cell zooming</td>
<td></td>
<td>interference</td>
<td>[79]</td>
</tr>
<tr>
<td></td>
<td>coordinated multi-point (CoMP)</td>
<td></td>
<td></td>
<td>[80], [81]</td>
</tr>
<tr>
<td></td>
<td>cell discontinuous TX/RX (DTX/DRX)</td>
<td></td>
<td></td>
<td>[82]–[84]</td>
</tr>
<tr>
<td></td>
<td>coordinated sleep and napping (CoNap)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the receiver and to make sure a sufficiently strong signal is received. A core semiconductor device of the RF PA is the transistor [86]. A PA circuit based on a field-effect transistor (FET) is shown in Fig. 1 for example. A voltage or current applied to one pair of the transistor’s terminals (gate and drain) changes the drain current flowing through another pair of terminals (source and drain) which is the amplified output. In other words, over fixed bandwidth, the PA input signal power $P_{in}$ (called alternating current (AC) RF-drive power) is amplified to $P_{out}$, which is the PA RF-output signal power. To do this, external direct current (DC)-input power $P_{DC}$ is supplied to the PA through a RF choke, where $P_{DC}$ is the main source of power consumption at the PA.

We will next describe the PA linearity and efficiency in the following two subsections. Various applications of PA are introduced in Appendix A.

A. PA Linearity

The practical elements of a transistor of PA, such as transconductance, drain capacitance, and gate capacitance, are nonlinear; therefore, the practical PA is nonlinear and the perfect linearity does not hold. The nonlinearity of PA produces harmonic and intermodulation distortions. The harmonic distortion is generated unintentionally at harmonic frequency, which is the integer multiplication of the single fundamental (input signal) frequency, and the intermodulation distortion is generated at any linear combination of multiple fundamental frequencies. Even though the harmonic and intermodulation distortions are well defined mathematically, the adjacent channel power/leakage ratio (ACPR) and the error vector magnitude (EVM) are more widely used to measure the nonlinear distortion in a practical wireless transmitter, in which strong linearity is required and complex digital modulated signals are involved. The ACPR is a ratio between adjacent channel’s total power (intermodulation signal) and main channel’s power (useful signal) to measure the out-band distortion, while the EVM is to assess the in-band distortion. Since the higher ACPR or EVM causes more significant performance degradation of detection at the receiver and SE degradation, an accurate model for the PA linearity is desired to design SE and/or EE efficient systems.

The PA linearity model at the transistor level is accurate but difficult to analyze. In contrast to the transistor-level PA model, a system-level PA model includes a few key parameters which are obtained from measurements, and it is typically tractable to analysis and reasonably accurate; therefore, it has been widely used to model PAs. The system-level PA model is divided into two types, either with or without a memory effect.

1) System-level Model with Memory Effect: Due to the capacitance and inductance in the circuits and the thermal fluctuation of the PAs, a frequency-domain fluctuation with memory arises in the transfer function of the PA. For example, Fig. 3 illustrates the effect of memory on distortion of a two-tone signal. Any nonconstant distortion behavior, such as amplitude or phase deviation of intermodulation (IM) responses, at different modulation frequencies (tones/subcarriers), is caused by memory effect [88]. As reported in [89], the memory effects are negligible when the system bandwidth is between $1\text{MHz}$ and $5\text{MHz}$. However, the electrical memory effects are severe for systems with wider than $5\text{MHz}$ bandwidth and the thermal memory effects are severe for systems using narrower than $1\text{MHz}$ signals. A few commonly used PA linearity models with memory effects are introduced as follows:

- **Volterra series model** employs the multivariate polynomial series to express the output signal as a function of the PA input signal, the memory length, and the delay time as follows [90]:

$$y(t) = \sum_{k=1}^{K} \sum_{m_1=0}^{M-1} \cdots \sum_{m_K=0}^{M-1} h_k(m_1, \ldots, m_k) \prod_{l=1}^{K} x(t - m_l \delta_l),$$  \hspace{1cm} (1)$$

where $h_k(\cdot)$ is the $k$th order Volterra coefficient that models the nonlinearity; $m_k$ is the delay; $\delta_l$ is the time delay; $M$ is the total number of delay taps, i.e., the finite length of memory; and $K$ is the degree of the polynomial, i.e., the nonlinearity order. A larger $M$ and $K$ with a smaller $\delta_l$ can improve the accuracy of the nonlinear model (e.g., $M = K = 11$ and $\delta_l = 1\mu s$ can capture the high nonlinearity of RF PAs in practice), yet the intensive computational complexity is intractable as the required number of coefficients increases exponentially with $M$ and $K$. Hence, the Volterra model is relevant if the...
nonlinearity is mild and for non real-time applications, as otherwise a truncation of the series yields poor modeling results.

- **Wiener, Hammerstein, Wiener-Hammerstein models** consist of two parts: a linear filter part A with memory (i.e., a linear time invariant system) and a nonlinear part B without memory (i.e., memoryless nonlinear system) [91]. The structures of Wiener, Hammerstein, Wiener-Hammerstein models are A–B, B–A, and A–B–A, respectively. For example, A–B means that the output signal is modeled by the input signal passing through A followed by B. These models can achieve relatively accurate modeling results with fewer parameters as compared to the Volterra model.

- **Memory polynomial model** reduces the required number of coefficients in a Volterra series model to \(M + 1\) by assuming the phases are independent [92]. This leads to an exponential reduction in the computational complexity, and consequently, we can employ the memory polynomial model to real time applications.

2) **System-level Model without Memory Effect**: Memoryless models basically assume that the previous PA output signal does not affect the current PA output signal. Amplitude-to-amplitude (AM/AM) distortion and amplitude-to-phase (AM/PM) distortion are used for the memoryless model. PM/AM and PM/PM distortions are typically ignored unless they are strong (e.g., a quadrature modulator with predistortion). The distortion components which are close to the carrier frequency are difficult to be filtered away, hence they are emphasized by using a passband model. However, to ease simulation and computation, a baseband PA model that represents the nonlinearity of complex baseband frequency approximation is more widely used than the passband model [93], [94]. The straightforward way to model the baseband PA is to use a polynomial function, yet to model the nonlinearity its required order increases significantly. Hence, the baseband model is typically used to represent the nonlinearity. A few commonly used, generic baseband models and PA-specific baseband models are introduced. The generic baseband models are very simplified PA models, which include ideal model, linearized model, and soft limiter model; thus, they provide analysis tractability regardless of PA types. However, the model may not be accurate. On the other hand, the PA-specific baseband models are used to achieve more accurate model of PA nonlinearity that depends on more parameters. In the survey, we briefly introduce three PA-specific baseband models including Rapp model, Saleh model, and Ghorbani model.

- **Ideal model** is a perfectly linear model of the PA. Specifically, the amplitude of input signal of PA is linearly amplified over all power regime, and at the same time, there is no distortion on the phase information after the amplification, i.e.,

\[
y = gx,
\]

where \(x\) and \(y\) are PA input and output signals, respectively, and \(g > 0\) is a linear gain. Herein, only linear gain \(g\) exists without phase distortion in the amplified signal. The ideal model is used for a reference baseline as shown in Fig. 4.

- **Linearized model** is the simplest model with nonlinearity [95]:

\[
y = gx + n,
\]

where \(n\) is the nonlinear distortion which is independent of \(v_{in}\) and modeled as a Gaussian noise based on Bussgang’s theorem [96]. The linearized model does...
not capture PA clipping effect at high power regime; therefore, it is applicable only for a system with high enough input back-off (IBO), which will be introduced in Section V.

- **Soft limiter model** is the simplest model that can capture the clipping effect [97], so it is widely used for tractable analysis. To represent the nonlinear distortion, an amplitude-dependent gain function $\Gamma(|x|)$ and a phase shift function $\Phi(|x|)$ are defined as

$$y = \Gamma(|x|) e^{2\pi j \Phi(|x|)}, \quad (4)$$

The soft limiter is modeled as follow:

$$\Gamma(|x|) = \begin{cases} |x|, & \text{if } |x| < v_{\text{sat}} \\ v_{\text{sat}}, & \text{otherwise} \end{cases}, \quad \Phi(|x|) = 0, \quad (5)$$

where $v_{\text{sat}}$ is the saturation voltage. The soft limiter model can capture the clipping effect in the high power regime. This model is applicable if the nonlinearity in the low power regime has been mitigated by applying linearization techniques, such as those introduced in Section V. Thus, the soft limiter model is appropriate for any types of PAs in system performance analysis. Note that the ideal model can be represented by $\Gamma(|x|) = |x|$ and $\Phi(|x|) = 0$.

- **Rapp model** is used to model the envelope characteristic of solid state PA (SSPA), especially, class-AB, with smoothness factor $p$ [98]:

$$\Gamma(|x|) = |x| \left(1 + \left(\frac{|x|}{v_{\text{sat}}} \right)^{2p} \right)^{-\frac{1}{2p}}, \quad \Phi(|x|) = 0. \quad (6)$$

Note that the soft limiter model is a special case of the Rapp model with $p = \infty$. This model may be applicable if the linearization techniques are not perfect (large $p$) or no such techniques is used (small $p$).

- **Saleh model** is used for modeling a traveling wave tube amplifier (TWTA) [99]:

$$\Gamma(|x|) = \frac{a_1|x|}{1 + a_2|x|^2}, \quad \Phi(|x|) = \frac{b_1|x|^2}{1 + b_2|x|^2}, \quad (7)$$

where $a_i$ and $b_i$ are the distortion coefficients. The Saleh model matches well the practical AM/AM and AM/PM characteristics of TWTA and it is useful as a basis for designing linearization techniques such as predistorters. However, the model is less accurate for SSPA near the saturation region [99].

- **Ghorbani model** is used for modeling a FET amplifier with small amplitude nonlinearity [100]:

$$\Gamma(|x|) = \frac{a_1|x^{a_2}|}{1 + a_3|x^{a_2}|} + a_4|x|, \quad \Phi(|x|) = \frac{b_1|x^{b_2}|}{1 + b_2|x^{b_2}|} + b_4|x|, \quad (8)$$

where $a_i$ and $b_i$ are the distortion coefficients.

---

In Fig. 4, AM/AM distortion characteristics of the baseband models in (2)–(8) are compared. Linearity of AM/AM of all baseband PA models is conserved well in the low power regime, yet it is distorted severely near the saturation power level.

### B. PA Efficiency

In addition to the linearity characteristics, PA efficiency is another critical parameter to derive the practical EE. The PA power consumption $P_{PA}$ depends on many other factors such as hardware implementation, DC power $P_{DC}$, load characteristics, and operating frequency. Majority of the power is, however, consumed by the DC power [14]. To obtain a tractable but accurate analysis, we approximate the PA power consumption as

$$P_{PA} \approx P_{DC}. \quad (9)$$

PA efficiency is then defined as the ratio between the PA output power and $P_{PA}$. More precisely, the overall efficiency is defined as [87]:

$$\eta_{\text{all}} \triangleq \frac{P_{\text{out}}}{P_{\text{PA}} + P_{\text{in}}}. \quad (10)$$

The overall efficiency $\eta_{\text{all}}$ is widely used as it is accurate over a broad range of PA gain $g$ [14]. Two other commonly used PA efficiency models with a high $g$ are drain efficiency $\eta_{\text{drain}}$ and power-added efficiency (PAE) $\eta_{\text{PAE}}$, which are defined as

$$\eta_{\text{drain}} \triangleq \frac{P_{\text{out}}}{P_{\text{PA}}} \quad \text{and} \quad \eta_{\text{PAE}} \triangleq \frac{P_{\text{out}}}{P_{\text{PA}}} - \eta_{\text{drain}} \left(1 - \frac{1}{g}\right). \quad (11)$$

Since the PAE includes the PA gain information $g$, it is a widely used metric in communications. In wireless communication systems, typically $\eta_{\text{all}} \approx \eta_{\text{drain}} \approx \eta_{\text{PAE}}$, because $P_{\text{in}}$ is relatively small compared to $P_{\text{DC}}$ and $P_{\text{out}}$. Ideally, the total
power consumption of PA, denoted by $P_{\text{PA}}$, equals the output transmit power $P_{\text{out}}$, i.e.,

$$P_{\text{PA}} = P_{\text{out}}$$

However, the maximum efficiency of real-life PA is typically between 20% and 35% [23], [26], as illustrated in Fig. 5 where the $(P_{\text{DC}}, P_{\text{max}})$ values of 115 PAs are obtained from data sheets. In other words, $\eta_{\text{PAE}} < 1$, and hence the ideal situation reflected by (13) does not hold in practice. We thus have

$$P_{\text{PA}} > P_{\text{out}}.$$  

On the other hand, a typical PA achieves its peak efficiency only at the peak envelop power (PEP) [85] and the efficiency drops rapidly as the output power decreases, as shown in Fig. 6 (the data is obtained from [26], [34], [101], [102]).

Before leaving this section, we emphasize the relationship between PA linearity and efficiency. As we see in Fig. 4, higher linearity can be achieved in the low power regime, while a typical PA achieves high efficiency in the high power regime as shown in Fig. 6. Therefore, there typically exists a fundamental tradeoff between the efficiency and the linearity [103]. Since the PA linearity can be sacrificed to increase the PA efficiency, and vice versa, the PA selection with proper type of class and transistor is important to design EE wireless communication systems. The choice of the proper PA depends on the system parameters and applications (refer to [37], [104]–[106] and Section IV-H).

## III. SE and EE Tradeoff

SE and EE are two fundamental figures of merit in wireless communications. Various types of EE metrics have been used for different applications, such as power usage efficiency, data center efficiency, telecommunications energy efficiency ratio (Gb/W), telecommunications equipment energy efficiency rating ($-\log(\text{Gb}/\text{W})$), energy consumption rating (W/Gb), area power consumption (Km²/W), user power consumption (users/W) [11], [107], [108], and network energy efficiency with delay effect (dB/J) [109]. Throughout the paper, we define EE as the total SE (b/s/Hz) over the allocated bandwidth per unit power, i.e., $b/J$. This EE definition has been typically also considered in many studies with various applications, such as the average of per-user EE [110], a network EE [111], [112], a practical network EE considering a packet error rate and information bits per packet [113], and a network EE considering a fixed target rate without a maximum transmit power constraint [114], [115]. The SE is defined as the maximum amount of bits that can be decoded reliably per unit time and per unit bandwidth [116].

### A. Theoretical SE-EE Tradeoff

The theoretical SE-EE tradeoff is given by [10], [117]:

$$\text{SE}(P_{\text{out}}) = \log_2(1 + P_{\text{out}}/\sigma^2),$$

$$\text{EE}(P_{\text{out}}) = \frac{\Omega \text{SE}(P_{\text{out}})}{P_{\text{out}}},$$

where $P_{\text{out}}$ is output transmit power; $\sigma^2$ is the noise power; and $\Omega$ is the total bandwidth used. The ideal SE($P_{\text{out}}$) in (15a) increases asymptotically as $\log_2(P_{\text{out}})$ as $P_{\text{out}}$ increases, and hence the ideal EE($P_{\text{out}}$) decreases asymptotically as $\log_2(P_{\text{out}})/P_{\text{out}}$. Thus, the SE-EE tradeoff region is a decreasing convex function as illustrated by dashed-line in Fig. 7, as also observed in [10], [117]. A few theoretical studies on the theoretical SE-EE tradeoff have been performed recently [11], [107], [108].

However, the theoretical SE-EE tradeoff in (15) is derived for an ideal PA that satisfies the following three assumptions:

A1. Perfect linearity of PA in (2).

A2. Ideal energy efficiency of PA in (13).

A3. No overhead power consumption: the total system power consumption $P_c$ is identical to the PA power consumption $P_{\text{PA}}$, i.e.,

$$P_c = P_{\text{PA}}.$$  

In practice, the assumptions in (2), (13), and (16) are invalid, resulting in a practical SE-EE tradeoff which has discrepancy from the theoretical SE-EE tradeoff in (15).
B. Practical SE and EE Tradeoff

Practically, as we pointed out in Section II-A, the PA nonlinearity yielding signal distortion at the transmitter makes (2) and the assumption A1 invalid. Hence, the ideal SE in (15a) is reduced to a practical SE, denoted by \( SE'(P_{out}) \), as
\[
SE'(P_{out}) < SE(P_{out}). \tag{17}
\]
Moreover, practical PA’s energy consumption models in (14) violates A2. Thus, we use (17) and (14) instead of A1 and A2, respectively, to derive practical SE-EE tradeoff.

In addition to (17) and (14), to improve A3 and to derive practical SE-EE tradeoff, we need a sufficiently accurate-and-tractable, but practical, transmitter’s power consumption model, which includes overhead power consumption. More concretely, the total power consumption is a function of \( P_{out} \) and modeled by two components [75]:
\[
P_c(P_{out}) = TPI(P_{out}) + TPD(P_{out}),
\]
where the first component TPI \( (P_{out}) \) is the transmit power independent (TPI) term and the second component TPD \( (P_{out}) \) is the transmit power dependent (TPD) term. The TPI includes, for example, power consumption for the alternating current to direct current (AC/DC) converter, DC/DC converter, and an active cooling system. On the other hand, the TPD term depends on the transmitter architecture, PA, and the applications. The TPD term is empirically modeled as a linear function of \( P_{out} \) as [38], [107], [118]
\[
TPD(P_{out}) = c_1 P_{out},
\]
where \( c_1 > 1 \) is a scaling constant. Alternatively, a nonlinear function can be used as [23], [26]
\[
TPD(P_{out}) = c_2 \sqrt{P_{out}},
\]
where \( c_2 > 1 \) is a coefficient depending on the PA type. Clearly, the total power consumption of the practical system is greater than the PA power consumption, i.e.,
\[
P_c > P_{PA}, \tag{18}
\]
thus making A3 invalid.

Considering the overhead circuit power consumption of (14) and (18), a more practical definition of EE is given by [109], [111], [113], [119]–[122]:
\[
EE'(P_{out}) = \frac{\Omega SE(P_{out})}{P_c(P_{out})}. \tag{19}
\]
However, the SE in (19) is still assumed to be ideal. In summary, contrary to the assumptions A1, A2, and A3, the practical systems have
\[
SE'(P_{out}) < SE(P_{out}) \text{ and } P_{out} < P_{PA} < P_c(P_{out}),
\]
and the theoretical SE-EE tradeoff in (15) is revised accordingly as follows:
\[
EE''(P_{out}) = \frac{\Omega SE'(P_{out})}{P_c(P_{out})}. \tag{20}
\]
Typically, \( SE'(P_{out}) \) is no longer a log function, and \( P_c \) is a function of \( P_{out} \). Recently, a semi-closed form expression for \( SE'(P_{out}) \) has been analytically derived and the practical SE-EE tradeoff in (20) has been obtained [23], [26]. From the practical models, asymptotically as \( P_c \) increases, the output power \( P_{out} \) saturates as shown in (4) and (5), and so SE saturates to some upper limit. Consequently, SE degradation is inevitable as illustrated by the horizontal double arrow in Fig. 7.

Due to the proportional relationship between SE and EE in (20), the SE degradation causes the EE degradation directly as illustrated by the vertical solid-line arrow in Fig. 7. In addition to the SE degradation from PA nonlinearity, an imperfect efficiency of system and PA causes overhead power consumption, resulting in severe EE degradation. From (15), for an ideal PA, we can achieve the EE value of \( \sqrt{\ln 2} \) as SE approaches zero. For the real-life PA, however, the overhead power/energy consumption makes the EE approach zero when the SE approaches zero.

In summary, the practical EE increases before a turning point and decreases sharply after the turning point. In other words, the practical SE-EE tradeoff has a range of Pareto-optimal tradeoff boundary that is very narrow. In modern wireless communications, however, a wide range of SE-EE tradeoff is desired due to the dynamic demand for high SE or EE depending on the service types, traffic, and channel conditions. In the remaining sections, we introduce various PA-centric energy efficient technologies, which can improve EE or SE-EE tradeoff.

IV. PA Design Approaches

In this section, we consider a PA design approach to improve EE for wireless communications. The PA design is based on constructing a transmit architecture that is a building block of various circuits, e.g., multiple PAs, oscillators, mixers, filters, matching networks, combiners, and circulators [14], [123]. Various transmitter architectures have been introduced as illustrated in Fig. 8. We concentrate on the PA linearity and efficiency in the structure to design an energy efficient wireless communications system.

A. Linear Architecture

Multiple cascade PAs are located in an amplifier chain, in which each amplification stage has around 6–20 dB gain. Refer to the example with three PAs in Fig. 8(a). To achieve high linearity, multiple class-A PAs each with large IBO are used before the last PA, i.e., for all of the driver stages. For the final PA, a class-B type is desirable to achieve high energy efficiency. With this design, the linear architecture achieves high linearity with modest efficiency.

B. Parallel Architecture

Multiple parallel PAs are used to achieve high reliability or high power amplification [56]. For reliable amplification, multiple isolated PAs in hybrid combiner are used in case some PAs fail, as depicted in Figs. 8(b1). Quadrature combiner is used to achieve a constant input impedance, to reduce the effect of a load impedance, and to cancel odd harmonics and
backward intermodulation distortion. For ease of fabrication, Wilkinson or in-phase power combiner is widely used [16].

For high power transmission, the power combining structure can be used as shown in Fig. 8(b2). The multiple PAs are simultaneously activated at their maximum power and the amplified signals are combined to improve the SE [17]. The decision on whether to employ a power combiner with multiple low power amplifiers (LPAs) or a single high power amplifier (HPA) should be performed carefully. Compared to a single HPA, multiple LPAs achieve higher gain, wider bandwidth, better phase linearity, and lower cost, while its assembly time and form factor cost increase significantly.

C. Switching Architecture

As mentioned in Section II-B, PA efficiency is maximized at PEP output, yet PA is typically operated below the PEP. In practice, the size of the transistor, quiescent current, and supply voltage are fixed to operate at the PEP, resulting in efficiency degradation when lower power transmission is used [123]. To circumvent the efficiency degradation, a stage-bypassing (or gate-switching) technique that switches/selects PAs with different maximum output power levels has been proposed [18]–[20]. The switching PA is performed based on the input signal power level, as illustrated in Fig. 8(c). Recently, a PA selection/switching (PAS) technique has been proposed for cellular networks based on the desired output signal power [21], [22], [25] and the channel state information (CSI) [23], [24], [26]. To achieve a wide Pareto-optimal SE-EE tradeoff region, the PAS transmitter selects one or more PAs at any time to maximize the EE while satisfying the required SE by using the CSI at the transmitter (CSIT) [23], [26]. For practical scenarios with PAS, a partial CSIT consisting of a feedback of the PA to be used for transmission has been considered [24]. Since each PA operates at its own maximum output power compared to the conventional power control systems, the overall efficiency can be improved by selecting the most efficient PA that can support a target rate with the least power consumption. For the PAS driven by a digital signal from a baseband signal processor (digital signal processor), a class-S modulators can be employed.

D. Envelope Tracking (ET) Architecture

Similar to the conventional linear RF PA, an envelope tracking (ET) PA amplifies an input signal containing both amplitude and phase information, refer to Fig. 8(d); however, its supply voltage is controlled by the envelope of DC voltage of the output of DC/DC converter, which is controlled by the input signal envelope, resulting in an energy efficiency that lies between the linear RF PAs and the PAs using the Kahn technique. The ET technique has been widely employed to improve PAE in various wireless transmitters [27]–[31].

E. Envelope Elimination and Restoration (EER) Technique

To achieve high efficiency and linearity for a wide range of signals and power (backoff) levels, the phase and envelop of the signal are amplified separately and then combined as shown in Fig. 8(e), which is thus referred to as the envelope elimination and restoration (EER) technique, or also as Kahn technique [32]. A highly efficient nonlinear RF PA, such as class-C, class-D, class-E, or class-F, amplifies the constant-amplitude phase-modulated (PM) signal while allowing for nonlinearity, while a highly efficient envelope amplifier, such as class-S or class-G, amplifies the envelope extracted from the input signal with high linearity, and then the final RF PA restores the envelope to the PM carrier. As a result, EER can achieve a high efficiency over a wide range of signal which is three to five times those of linear amplifiers from HF band to L band.

F. Outphasing Technique

Outphasing technique performs linear amplification using nonlinear components (LINK) [33]. To achieve high linearity of amplitude-modulated (AM) signal, the AM signal is split into two constant-envelope PM signals $P_1$ and $P_2$ by a signal component separator (SCS) in Fig. 8(f). Since the signal splitter and power combiner do not introduce nonlinear distortion, and there is no information of the magnitude of PM signals, high linearity and efficiency can be achieved by amplifying the two PM signals individually through the nonlinear yet high-efficiency PAs. After the separate nonlinear amplification, two amplified PM signals are combined to regenerate an amplified replica of the original AM signal.
G. Doherty Technique

To improve efficiency, two different types of PAs, namely class-B (carrier PA or main PA) and class-C (peaking PA or auxiliary PA) PAs, are combined as shown in Fig. 8(g). The class-B PA is activated if the signal amplitude is half or less than the PEP, while both PAs are activated if the signal amplitude is larger than half of the PEP [124]. Contrary to EER, ET, and outphasing techniques, which requires an external control circuits and signal processing, Doherty technique needs only a quarter-wave transmission line to combine two output powers, and thus it is attractive for implementation [35]. Furthermore, due to the high efficiency and linearity over wide frequency band and power level, Doherty technique is widely used for modern cellular communications [1], [34], [36]–[39], along with EER, ET, and outphasing architectures.

H. Design Applications

A high linearity PA structure focuses on increasing the numerator of (20), while a high efficiency PA structure focuses on decreasing the denominator of (20), resulting in the improvement of the practical EE. As pointed out in Section III, however, there typically exists a tradeoff between the linearity and the efficiency of the PA; therefore, we need to carefully design/determine the proper PA for high EE. As mentioned in Section II, PA can be chosen based on system parameters and applications. In particular, we discuss PA designs depending on two important parameters: communication coverage and bandwidth.

Coverage: A short-range (e.g., WiFi and Bluetooth) PA has typically small power output capability (POC) and is implemented with small form factor at low fabrication cost; therefore, sophisticated techniques to improve the linearity (refer to Section V) may not be available. Hence, a low complexity techniques is employed to circumvent the high nonlinearity, such as high employing IBO at around 12dB, resulting in typically low efficiency at around 8% [125]. On the other hand, powerful linearization techniques can be implemented in practice for a long-range (e.g., cellular base station) PA with less constraint on the manufacturing cost and form factor. Thus, for a long-range PA, high power transmission with less IBO is possible, and the PA efficiency is typically higher (about 30% – 80%) than the short-range PAs.

Bandwidth: Cognitive radio (CR) technology is a key solution to reduce the inefficiencies in the usage of the spectrum [126]. CR technology enables a secondary transmitter to access the spectrum that is licensed to a primary user, in a dynamic manner through spectrum sensing. Two issues arise in the CR with respect to the EE. The first issue is the high sensing energy consumption. In [127], an EE is considered for a relay node that help for the secondary user to sense the primary user signal in CR networks. Since the high sensing performance increases a network SE indirectly, there exists a tradeoff between SE and EE. The second issue is on broadband signalling. Since the CR transmitter access the multiple frequency band, PA needs to have broadband capability. Several PA techniques, such as multi-stage technique and push-pull PA, have been developed to provide high PAE for broadband frequency band [128] and references therein. In [128], a load-tracking technique is proposed for developing wideband PA. In [129], a multi-level and multi-band PA is proposed for CR. Due to its multi-level architecture, the proposed PA architecture provides peak efficiency at multiple output power. The multi-band operation is enabled by tuning a resonator.

I. Summary and Discussion

In this section, we introduced a transmit architecture for energy efficient communications systems, such as linear, parallel, switching, and envelope tracking architectures and envelope elimination and restoration, outphasing, Doherty techniques. For further details on design of energy efficient PA, refer to Appendix A and references [14], [15], [56], [86], [87], [123], [130]–[132]. As surveyed in this section, it is difficult to obtain exact and practical PA models. Therefore, rather than a mathematical optimization approach, the typical way to design/optimize PA is based on modeling and empirical evaluation, i.e., a simulated-annealing-based current computer-aided design tool [133]–[137]. However, nonlinear and inefficient (η < 1) PA characteristics appear to be fundamental and inevitable. In the next sections, we further survey signal and network design based technologies to enhance EE against the fundamental limitation of the PAs, for which optimization techniques can be applied successfully.

V. Signal Design Approaches

The PA architectures presented in Section IV improve the PA nonlinearity and the PA efficiency, yet the improvement is limited and still unideal. We exploit knowledge of the signals to mitigate the adverse impact of PA nonlinearity and imperfect PA efficiency through signal processing and coding techniques, and eventually to further improve EE. In this section, we first survey six classes of PAPR reduction techniques to mitigate the severe distortion introduced when the signal amplitude exceeds the linear amplification region of PA [138]. In the moderate output power regime, the output of the PA displays relatively smaller nonlinear distortion. Next, we categorize four linearization and predistortion techniques to mitigate the distortion. Finally, we discuss the advantages and disadvantages of the different techniques.

A. PAPR Reduction

To avoid the occurrence of high signal peak, various PAPR reduction techniques have been proposed, namely, clipping, coding, partial transmit sequence (PTS), selective mapping (SLM), tone reservation (TR) and tone injection (TI) [139]. If the PAPR is reduced, the average transmit power can be increased. Therefore, the SE increases. However, as it will be described in the following, each PAPR reduction technique has a trade-off between the PAPR reduction and the SE loss and the increased complexity.
1) **Clipping** [40]–[46]: The PEP is restricted to a predetermined maximum clipping level. The clipped signal with saturation volatage of $v_{\text{sat}}$ is given by (5). The complexity of clipping is independent of the number of subcarriers. This simplicity is the key advantage of clipping over other techniques. However, the clipping operation is nonlinear which incurs SE degradation [26]. The SE degradation can be compensated to some extent by channel coding [43] and signal processing at receiver [44], [46]. The nonlinear clipping operation also causes leakage into adjacent channels, i.e., out-of-band (OOB) radiation. The repeated use of clipping and filtering can effectively mitigate this OOB radiation at the expense of the computational complexity [45].

2) **Coding** [47], [48], [140]–[143]: The original binary information sequence is mapped onto longer codeword sequences by coding, such as block coding [47] and Golay complementary sequences [48], [140], [141]. An example of a simple block coding scheme [47] is to map three bits to four bits by adding one parity check bit. It is known that the Golay complementary sequences can be used as a codeword to reduce the PAPR [140]. Recent studies have shown that 16-QAM sequence with low PAPR can be constructed from Golay sequences [141], [142]. In [143], a code based on dual Base-Ray-Chaudhuri (BCH) codes, which achieves PAPR reduction by 7dB, is proposed. Some coding schemes require side information to inform the location of the additional bits for PAPR reduction. However, since the coding introduces redundancy, the SE would be reduced.

3) **Partial Transmit Sequence (PTS)** [51], [144], [145]: The frequency domain data block to be transmitted is partitioned into disjoint subblocks. Each subblock is multiplied by a rotation factor. The optimum combination of the rotation factors, which minimizes PAPR, is then conveyed to the receiver as side information. To reduce the complexity required for searching the optimum combination of the rotation factors, several algorithms have been proposed, e.g., iterative flipping algorithm [144] and neighborhood gradient-descent search [145]. The number of subblocks, the number of rotation factors, and the input data block partitioning should be optimized for sufficient PAPR reduction [51]. However, the joint optimization increases complexity at the transmitter [139], and the huge side information including rotation factors degrades the SE.

4) **Selective Mapping (SLM)** [52], [53], [146]: Independent transmit sequences representing the same information are generated at the transmitter. The sequence with the lowest PAPR is then selected for transmission. In [52], each sequence is generated by multiplying the original information sequence by independent phase rotation sequences. To recover the original information sequence, the receiver needs to know the selected phase rotation sequence. Therefore, a dedicated side information is necessary. To avoid sending the side information, several blind algorithms were proposed [53], [146]. In [53], an SLM with scrambling, which requires no explicit side information, is proposed. The maximum likelihood decoder is proposed for a blind SLM proposed [146], which provides the same performance as the SLM with perfect recovery of the side information. Multiple operations of inverse fast Fourier transform (IFFT) process are required at the transmitter for an SLM. As such, the computational complexity for encoding at the transmitter and decoding at the receiver could be prohibitive for real-time implementation.

5) **Tone Reservation (TR)** [54]: A small subset of subcarriers is reserved for transmitting special signals to reduce PAPR. The optimization of the location of the reserved subcarriers and power allocation is necessary to maximize the PAPR reduction. For example, the optimum special signal can be obtained via a linear programming (LP) problem whose computational complexity is prohibitively high. To reduce the complexity, a gradient algorithm is proposed in [54]. Since the reserved subcarriers can not be used for information transmission, there is SE degradation.

6) **Tone Injection (TI)** [54], [147]: The original constellation point is mapped onto one of several possible constellation points with larger constellation size. The additional degree of freedom due to the larger constellation size can be used for PAPR reduction. Different from other techniques, TI does not require information exchange between the transmitter and the receiver. Although TI significantly reduce the PAPR, additional receiver processing is necessary to recover the original constellation points. Furthermore, the average signal power increases due to its larger constellation size [139]. A special structure of constellation, such as a hexagonal constellation in [147], can be used to reduce the average transmit power.

### B. Linearization/Predistortion

Fig. 4 illustrates that a significant distortion is introduced if the signal amplitude exceeds a threshold. Therefore, the IBO and PAPR reduction techniques mainly focus on preventing/avoiding the signal from having such a high PEP. However, a nonlinear effect causing distortion exists even in the low output power regime. To compensate this nonlinearity effect, various linearization methods have been rigorously studied, which are classified as feedforward, feedback, and digital predistortion. Although additional analog circuits and computational complexity are required, these linearization techniques can improve SE by compensating the distortion introduced by PA to the signal.

1) **Feedforward Linearization** [55], [56]: Feedforward linearization is implemented in analog domain. First, the PA input signal is subtracted from the PA output signal in order to obtain the distortion introduced by the PA. Then, the distortion is amplified by a separate auxiliary PA. Finally, the amplified distortion is subtracted from the original PA output signal. The feedforward linearizer is nonparametric, i.e., it does not rely on any priori knowledge of signal and PA characteristics. However, analog feedforward approach is complex and sensitive to component tolerance and drift and to the change in input power level. Furthermore an auxiliary PA with good linearity is required to linearly amplify the distortion, which will be subtracted from the original PA output signal. These issues make practical implementation of feedforward linearization challenging [55].

2) **Feedback Linearization/Analog Predistortion** [57]–[61]: The PA nonlinearity dynamically changes due to aging and
the operating frequency. Since DPD is performed in digital domain. On the other hand, DPD is independent of the varying operating temperature. The dynamically changing PA nonlinearity introduces dynamic distortion to the signal. The distortion is compensated by a closed-loop configuration in feedback linearization. Cartesian feedback is often used for feedback linearization [61]. The feedback linearization mainly compensates for the third order component introduced by the PA and can provide a good performance without high implementation complexity. Since no distortion is introduced to the signals before the predistorter, which is located before a PA, the same clock speed with the original clock speed is sufficient for intermediate frequency (IF) filter bandwidth to be sustained by the original signal bandwidth. Note that higher clock speed is required to capture the signals that is distorted and expanded to broader bandwidth. The use of a low, common clock speed greatly relaxes the IF filter bandwidth and hence leads to a low power consumption and a simplified architecture [130]. However, to implement feedback linearization, a considerable amount of additional analog hardware is required [65].

3) Digital Predistortion (DPD) [62]–[65]: The feedforward/feedback linearization (analog predistortion) techniques depends on the operation frequency as they are implemented in analog domain. On the other hand, DPD is independent of the operating frequency. Since DPD is performed in digital domain, it can compensate PA's irregular characteristics and also accurately compensate the distorted signal over a wide dynamic range. An adaptive algorithm can be used to select the correction factor stored in look-up table by comparing the digital signal and the feedback information [148]. Further linearization performance improvement can be achieved by combining DPD and PAPR reduction techniques [149]. The predistortion introduced in the digital domain expands the bandwidth. Therefore, the faster clock speed and the wider IF filter bandwidth than the original signal are required, with additional feedback and significant signal processing, which increases system power consumption [148].

C. Discussion

The signal processing techniques introduced in this section can avoid/compensate the adverse impact of the PA nonlinearity. However, the benefit of signal linearization can be obtained with cost, such as PA efficiency reduction (high power consumption) for IBO, OOB radiation increase for signal clipping, SE reduction for side information, additional computational complexity (see e.g., parameter optimization and multiple candidate sequence generation [139]), and additional analog circuits. Selecting appropriate PAPR reduction technique and linearization technique is important to satisfy performance requirement and hardware capability. For example, sophisticated technique such as DPD could be used in a BS but not in a user terminal due to its large size and high power consumption.

VI. NETWORK DESIGN APPROACHES

PA nonlinearity can be mitigated by the signal processing techniques presented in Section V. However, the system still suffers from imperfect PA efficiency. The adverse impact of the imperfect PA efficiency becomes significant when the transmit power is far less than PA's maximum operating power level, i.e., PEP. Herein, we focus mainly on improving the EE of macro BSs with cell coverage of several hundred meters to several kilometers [150], which can magnify the EE benefit via the network perspective. This is because 72% of the total network energy is consumed by the BSs [152], where the transmit power of the HPAs ranges from 43dB to 46dB. In this section, we review the network densification and the network protocol that can reduce the adverse impact of HPAs' imperfect efficiency.

A. Network Densification

To provide reliable communication to the geographically distributed users without drastic increase of transmit power, network densification that shortens the transmission distance between the transmitters and the receivers is the most effective way. This approach tries to avoid the use of a HPA by deploying more transmit elements in the network. To support the ever growing traffic demand with low power consumption is very challenging for wireless communications. A heterogeneous network (HetNet) architecture [66], [67] is a promising solution to improve the network’s EE. The HetNet particularly aims network EE improvement for high-traffic-load network condition. Based on various criteria, such as backbone network architecture and deployment scenario, HetNet can be classified into several broad categories (refer to the illustration in Fig. 9): multiple radio access technologies (RATs)/small cells [66], [67] (they can be further classified into in-band and out-band small cells based on used spectrum), relays with a low power amplifier (LPA) [76], cooperative relaying [5], [153], and distributed antenna system (DAS) [72]–[75], [154], [155].

1) In/Out Band Small Cell: New small BSs are deployed within macro BS coverage as shown in Fig. 9. A part of the traffic at a macro BS is offloaded to those nodes, i.e., traffic offloading [68]–[71]. A macro BS provides low-/medium-rate service to a large area, while small BSs provide high-rate service in so-called hotspot area. The shorter transmission distance between a small BS and a user equipment (UE) significantly can reduce the required transmit power for reliable communication. Therefore, an LPA can be used instead of a HPA. The typical power consumption of a small BS is much lower than that of a macro BS [125], and hence the network’s EE significantly improves [156]. Recently, the deployment of a large number of antenna elements at a BS, called large-scale/massive multiple-input multiple-output (L-MIMO or massive MIMO), has been gaining much attention [157], [158]. In [159], the concept of soft-cell cooperation is proposed for small BSs and a macro BS with massive MIMO to greatly reduce the total power consumption, i.e., improve the network’s EE. Depending on the spectrum use, we can deploy two types of small cells as follows:

- **In-band small cells**: A pico cell and a femto cell that adopt the same RAT as the macro BSs [66] can be

3In wireless sensor network (WSN), the key objective is to prolong the battery life of sensors while satisfying the system requirement. In WSN, the sensor node is equipped with low-power PA as its transmit power is quite small. Therefore, the network’s EE of WSN can be improved by energy-efficient routing techniques [151].
considered as in-band small cells. The transmit power of a macro BS is several orders of magnitude higher than that of a small BS. Therefore, a UE associated with small BSs is subjected to huge interference from a macro BS, i.e., cross-tier interference from macro BSs. To reduce the cross-tier interference, enhanced inter-cell interference coordination (eICIC) has been considered, in which orthogonal resources are assigned to macro and small BSs [160]. On the other hand, if a number of in-band small cells are deployed within macro cell coverage, cross-tier interference from small cells to macro UEs who are associated with macro cell becomes significant. In [161], an energy efficient cross-layer radio resource management is proposed for (in-band) small cells. To mitigate the cross-tier interference from femto BSs to macro UE along with intra-tier interference from macro BSs to the macro UE, a distributed power adaptation algorithm is developed to satisfy QoS requirement of UEs. To reduce the interference among different tiers, the coordination and the optimization of resource utilization are essential. This may increase the complexity of the system.

- **Out-band small cells**: We call small cells that operate in different RAT and frequency from existing macro BSs as out-band small cells. This concept encompasses WiFi offloading [70] and phantom cell/soft cell [162]. Different from in-band small cell, there is no cross-tier interference as the out-band small cells use different/orthogonal frequency band with much wider bandwidth. Hence, high data transmission is achieved.

2) **Distributed Antenna System (DAS)**: DAS consists of multiple antenna ports (APs), or so-called a remote radio head [72]–[75], [154], [155]. The multiple APs are geographically distributed within the current macro BS coverage as shown in Fig. 9. The multiple APs are connected to a macro BS which is acting as a central unit (CU) via a dedicated wired backhaul link (e.g., optical fibre). This effectively shortens the transmission distance, and hence avoids the macro BS from high power transmission. The signal processing at CU enables the APs to jointly process the transmitted or received signal. In [75], an optimization problem is formulated to maximize the network EE in large-scale DAS. The original EE maximization problem is split into subproblems to obtain the precoding weight and power control. Although DAS is a promising approach to improve EE, the additional circuits and processing by a network protocol. This approach tries to reduce the PA active duration, so that overhead power consumption can be mitigated. The PA power consumption at each transmission time instance is determined by the frequency domain resource utilization as illustrated in Fig. 10. During peak-traffic period, i.e., the first slot in Fig. 10(a), a larger fraction of resources will be allocated through a macro BS to the macro UE along with intra-tier interference from macro BSs. To reduce the cross-tier interference from small cells to macro UEs, a distributed power adaptation algorithm is developed to satisfy QoS requirement of UEs. To fully utilize the benefit of DAS, an optimization of beamforming weight and user scheduling are necessary [75]. This may increase the complexity of the system.

3) **Relay**: New nodes with LPA controlled by a macro BS are deployed. A relay forwards the signal from a macro BS to a UE [5], [76]. The deployment of relay with LPA has a few advantages. The macro BS coverage can be expanded without increasing its transmit power. Therefore, smaller number of BSs are sufficient to support the coverage. This reduces unnecessary power consumption. The macro BS coverage can also be served by lower transmit power of a macro BS [77]. Furthermore, the macro BS coverage can be expanded with negligible increase of power consumption. However, since the backhaul link between the macro BS and relay is typically implemented by wireless links, the dedicated radio resources, such as time slots and frequency bands, are necessary, which may offset the benefit of relaying.

**B. Network Protocol**

To reduce the overall overhead power consumption, the PA can be turned on and off dynamically via cooperation and processing by a network protocol. This approach tries to reduce the PA active duration, so that overhead power consumption can be mitigated. The PA power consumption at each transmission time instance is determined by the frequency domain resource utilization as illustrated in Fig. 10. During peak-traffic period, i.e., the first slot in Fig. 10(a), a larger fraction of resources will be allocated through a macro BS to the macro UE along with intra-tier interference from macro BSs. To reduce the cross-tier interference from small cells to macro UEs, a distributed power adaptation algorithm is developed to satisfy QoS requirement of UEs. This may increase the complexity of the system.

4) Typical PAAs achieve relatively high efficiency only within limited output power, i.e., only when the power is close to its maximum output power. The network protocol schedules the PA to operate mostly in high PA efficiency regime to improve network’s EE by exploiting the different EE at different output power.
proper scheduling and network processing including cooperation among the nodes. Hence, the activated HPA operates at close to its maximum efficiency. On the other hand, during the non-peak traffic period, i.e., the third and the fourth slots in Fig. 10, a smaller fraction of the resources is sufficient to satisfy the QoS requirement; therefore, the frequency domain resource utilization would be low and the PA operates with low PA efficiency. To enhance the EE, the scheduler manages resource allocation, so that the transmitter directly/indirectly operates the PA at close to its maximum efficiency and turns off it during the inactive period. Contrary to the network densification that can improve EE under high-traffic-load network condition, the network protocol approach can improve the network EE under low-traffic-load network condition. The network protocol can realize activation and deactivation with micro-scale control, such as discontinuous transmission (DTX) [80] and coordinated sleep/napping (CoNap) [82]–[84], or macro-scale control such as cell zooming [78], coordinated multi-point (CoMP) transmission [121], [163].

1) **Cell discontinuous transmission (DTX):** Cell DTX allows a BS to squeeze its traffic into small number of subframes as shown in Fig 10(b). During those active subframes, the frequency domain resource utilization becomes high, therefore PA operates close to its maximum efficiency. During other blank subframes, the HPA is put into inactive mode [80]. The fractional use of frequency resource results in the situation that PA operates with low PA efficiency. By aggregating the traffic into few active time slots, the PA output power is close to its maximum possible power, and hence achieves a high PA efficiency. This relaxes the requirement for PA to have relatively high PA efficiency over a broad operation range. Due to the high PA efficiency during active subframes and the deactivation of PA during blank subframe, significant amount of energy saving is achieved. This approach requires only a slight modification of evolved NodeB (eNB). Therefore it is considered as an attractive enabler of energy saving in 3rd generation partnership project (3GPP) [81]. Each BS independently performs cell DTX based on its associated users and their QoS requirement. The lack of coordination of blank subframe across different BSs may limit its energy saving gain.

2) **Coordinated sleep/coordinated napping (CoNap):** These approaches take into account the interference among neighboring cells. Similar to Cell DTX, the traffic is squeezed into subframes. However, different from cell DTX, CoNap squeezes the traffic by considering the interference from/to neighboring cells. The benefit of coordinated sleep among neighboring cells is elucidated in [82], [83]. It reduces the inter-cell interference (ICI). To further enhance EE improvement with time-varying traffic demand, adaptive CoNap is proposed in [84] that adaptively changes the napping pattern. In [84], it is shown the energy saving gain is obtained by clustering neighboring BSs and performing the CoNap. The coordination in the cluster requires information exchange among BSs. These methods require real-time monitoring of the traffic load at the networks. Nevertheless, as the CoNap does not require any sophisticated optimization, the CoNap can be implemented easily compared to the CoMP, which will be introduce subsequently.

3) **Coordinated Multipoint (CoMP):** CoMP coordinates the transmission from multiple active BSs to support the UEs in the dormant area as shown in Fig. 11(b) [79]. Information exchange among active BSs may necessitate a high speed backhaul [164]. Recently, energy-efficient resource allocation is proposed under limited backhaul capacity constraint [121]. The EE maximization optimization problem is solved via an efficient iterative resource allocation. In [163], a joint network optimization and downlink beamforming is proposed to minimize the overall BS power consumption while satisfying QoS requirements. Although the optimization of beamforming weight provides significant EE improvement, the incurred additional computational complexity may not be negligible.

4) **Cell Zooming:** Macro BSs are deactivated for up to several hours during non-peak traffic load time as shown in Fig. 11(c). The coverage area of the deactivated BS, which we call the dormant area, needs to be supported by other active BSs to sustain communications. Cell zooming expands the active BS coverage to support the UEs in the dormant area by the transmit power increase or the antenna tilt angle.
adjustment [78]. Cell zooming can be implemented with negligible complexity increase. To zoom out/in the coverage, the network controller monitors the traffic load of the system. Based on the monitored traffic load, BSs can be deactivated or activated. Since larger transmit power may be needed for cell zooming, the energy saving due to the BS deactivation could be discounted. Furthermore, the higher transmit power in cell zooming needs to be enabled by a HPA.

C. Discussion

The network densification, e.g., HetNet, brings significant improvement in network’s EE as discussed in this section. In HetNet, however, the cross-tier interference between different types of nodes needs to be carefully managed. In 3GPP standardization, a number of solutions have been considered, such as cell range expansion (CRE) [66] and eICIC [160]. The optimization of beamforming and radio resource management [165], [166] can further enhance the EE brought by the network densification.

The network protocol dynamically turns on or off a PA according to time-varying traffic demand to bring power saving gain. To this end, the cooperation or the coordination among different nodes is essential. Herein, it is important to decide how often the cooperation/coordination performs PA on and off to balance the system performance and the power saving. Through an optimization of the radio resource allocation, the opportunity for PA to be turned off can be increased. Therefore, further power saving can be achieved.

VII. CHALLENGES AND FUTURE OUTLOOK

In this paper, we have surveyed linearity and efficiency models of practical power amplifiers used in wireless communication systems and introduced various PA-centric technologies to improve energy efficiency for wireless communications. The PA-centric technologies have been categorized based on three design perspectives, namely, PA design, signal design, and network design. Some of the key challenges for reducing energy via these perspectives have been as follows (refer to Table 1):

- **PA design**: Ideal PA design in terms of linearity and efficiency is a fundamental solution for energy-efficient communications, yet, at the same time, there are fundamental difficulties, such as high manufacturing cost and large form factor. We omit studies on improving PA performance using advanced materials and metamaterials, which is out of scope of this survey.
- **Signal design**: Other issues, such as additional resource requirement, high sensitivity to PA status, and additional circuits, also restrict the immediate employment of the methods.
- **Network design**: On the other hand, network design methods may require high infrastructure cost, high overhead signalling, and frequent handover. Furthermore, scalability to already-deployed networks may arise as a new issue.

Moving forward, wireless communications will become even more ubiquitous. In addition to the current applications, such as human-to-human multimedia communications, machine-to-machine, human-to-machine and machine-to-human communications have been predicted to grow tremendously, connecting billions of devices. The applications are therefore diverse, the same as the QoS for the various applications [167]. For example, new personalized services with low latency, such as for intelligent transport system [168] or healthcare [169], has been receiving more attention for next-generation networks. Given this new trend, the infrastructure (e.g., base station, access point, and networks) will have to support these heterogeneous applications with a very likely heterogeneous network. The devices, especially the personalized devices, will have to support the heterogeneous applications in the various different application scenarios, hence posing a lot more design challenges to achieve energy efficient quality of service-guaranteed communications. [66], [67].

As such, the design of PA-centric energy-efficient technologies may need to take into account additional design metric. From the PA design and signal design perspectives, instead of focusing only on metrics such as high power efficiency, low peak-to-average power ratio, and bits per energy, i.e., energy efficiency, (or energy per bit), the system design may have to consider jointly with new metrics for specific applications, such as taking into account latency or deadline requirement [170]. The support of heterogeneous data over heterogeneous networks with varying power requirement and duty cycle, such as an access point that concurrently supports communication to wireless meters, wireless local area networks (LANs), and to cellular base stations, presents many interesting opportunities for energy saving. For example, an energy-aware PA may adopt a flexible architecture to support different wireless connections and power requirement.

Even more so, we believe many challenges, and hence opportunities, arise from the network design perspective. Future communication network will likely be more tightly coupled physically due to the increase in communication node density. In non-cooperating networks, the study of interference mitigation and game theoretic approaches may be relevant. The tight coupling of network elements may also be due to deliberate design, when different communication nodes cooperate or coordinate their transmission via a distributed antenna system or by a fronthaul infrastructure. In this case, acquisition and sharing of the system-state and network-state information are also critical issues that require further investigation due to the increase in energy needed.

APPENDIX

**CLASSES AND APPLICATIONS OF PA**

Depending on the circuit configuration, various classes of PA exist with different linearity and efficiency characteristics. For example [14], [15].

- class-A PA has very high linearity, yet the efficiency $\eta$ is lower than 50% (or power-output capability), and is generally used for millimeter wave (20-100GHz)
Fig. 12. Applications of power amplifiers over IEEE frequency bands. Note that the boundaries of PA types can be changed according to silicon semiconductor technologies and the applications.

- class-B PA has moderate linearity with $\eta \leq 78.5\%$ and is generally used for broadband at high frequency (HF) and very high frequency (VHF); class-AB has high linearity with $\eta \leq 78.5\%$
- class-C PA has low linearity with $78.5\% \leq \eta \leq 85\%$ and is used for high-power vacuum-tube transmitter
- class-D has very low linearity, yet very high efficiency and is used for 100W to 1kW transmitter at HF; class-E has low linearity, yet high efficiency at K-band (19 – 26.5GHz)
- class-F has low linearity, yet high efficiency at ultra HF (UHF) and microwave.
- class-G modulator is a combination of linear series-pass (class-B) PAs to reduce supply voltage.
- A class-S modulator uses a pair of transistors or a transistor with diode to generate a rectangular waveform whose width is proportional to the desired output signal amplitude. Since the class-S PA has high efficiency over wide dynamic range (ideally 100%), it is typically used as part of a Kahn-technique transmitter which will be introduced in Section IV [123].

The Si LDMOS PA are employed for many communication equipments. Especially, the international technology roadmap for semiconductors (ITRS) reports that 48 volt LDMOS transistor is widely used in cellular infrastructure market in 2011 [39]. This is because the LDMOS can support high output power required for a BS in cellular networks. To compensate poor efficiency of LDMOS devices in the low power regime, envelope tracking architecture or Doherty technique (see Section IV for the details) is used for 3G and 4G networks that employs high PAPR waveform signals.

**REFERENCES**


Sumei Sun (SM’12) received the B.Sc. (with honors) degree from Peking University, Beijing, China; the M.Eng. degree from Nanyang Technological University, Singapore; and the Ph.D. degree from National University of Singapore, Singapore.

She has been with Institute for Infocomm Research (I²R), Agency for Science, Technology, and Research (A*STAR), Singapore, since 1995, where she is currently Head of the Advanced Communication Technology Department, developing energy- and spectrum-efficient technologies for the next-generation communication systems. Her recent research interests include 5G transmission technologies, renewable energy management and cooperation in wireless systems and networks, and wireless transceiver design.

Dr. Sun served as Track Co-Chair of Mobile Networks, Applications, Services, IEEE Vehicular Technology Conference (VTC) 2014 Spring, Track Co-Chair of Transmission Technologies, IEEE VTC 2012 Spring, TPC Co-Chair of 14th (2014) and TPC Chair of 12th (2010) IEEE International Conference on Communications, General Co-Chair of 7th (2010) and 8th (2011) IEEE Vehicular Technology Society Asia Pacific Wireless Communications Symposium (APWCS), and Track Chair of Signal Processing for Communications, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2010 (APSIPA ASC 2010). She is also an Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and Editor for IEEE WIRELESS COMMUNICATION LETTERS. She receives the “Top Associate Editor” recognition in 2012 and 2013, and “Top15 Outstanding Editors” recognition in 2014, all from IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. She is a distinguished lecturer of IEEE Vehicular Society 2014-2016, a co-recipient of the 16th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications Best Paper Award, and Distinguished Visiting Fellow of the Royal Academy of Engineering, UK, in 2014.