5G-NR Bandwidth Efficient Modulation Options for Efficient Link Operation that are Compatible with mmW Transistor Nonlinearities

McCune, E.

DOI
10.1109/5GWF.2018.8516944

Publication date
2018

Document Version
Final published version

Published in
IEEE 5G World Forum, 5GWF 2018 - Conference Proceedings

5G-NR Bandwidth Efficient Modulation Options for Efficient Link Operation that are Compatible with mmW Transistor Nonlinearities

E. McCune; TU Delft

***Abstract* – The 5G network has a goal to significantly increase its energy efficiency with respect to the energy draw of the LTE network. The decision to use OFDM-based modulation for 5G-NR forces a low-valued ceiling on achievable energy efficiency from any linear power amplifier. This ceiling is lowest for amplifiers operating at frequencies near and above 30 GHz: the millimeter-wave bands. Transistors behave differently at these high frequencies, and the modulation used must change to match to these different characteristics. Ring oriented constellations with polar filtering meet these changed conditions.**

Keywords: 5G, new radio, transmitter, efficiency, millimeter wave

I. INTRODUCTION

Within technical communities that are not involved with implementation of standards in hardware, there appears to be an implicit assumption that modulation type and hardware performance using that signal are mutually independent. For example, the radio access network (RAN) committees within 3GPP have adopted modulations for Fifth Generation New Radio (5G-NR) that are identical for all operating bands, from those at 6 GHz and below, and for the millimeter wave bands at 27 GHz and above.

This assumption is not valid, because the transistor types available to radio hardware designers do behave differently at millimeter waved (mmW) frequencies. Further, signal modulation must also change to match implementation capabilities and meet 5G system objectives.

As product developments begin on 5G-NR transmitters, experience to-date already shows that low PA efficiency is a problem, especially at mmW frequencies [1] [2]. Reported PA efficiencies are nearly all below 10%, with some in the low single digits. This efficiency problem is predictable when a linear power amplifier (PA) is used [3]. Fortunately this prediction is constructive, in that insight into how circuit and signal construction can be done so that energy efficiency is greatly increased while maintaining bandwidth efficiency. Following these guidelines provides the design described in this work.

Energy efficiency is economically important for the reasons shown in Fig. 1 [4]. When circuit efficiency is below 40%, then the cost and size of the power supply and associated heatsink grow rapidly, with no corresponding increase in communication effectiveness. Once the circuit efficiency falls below 10%, the power supply and heatsink dominate the size (and cost) of the product. In effect, the PA and its associated circuitry become a heat pump, transferring most of the product supply-power into heat in the heatsink. Any useful wireless communication is effectively a small byproduct with respect to energy costs.

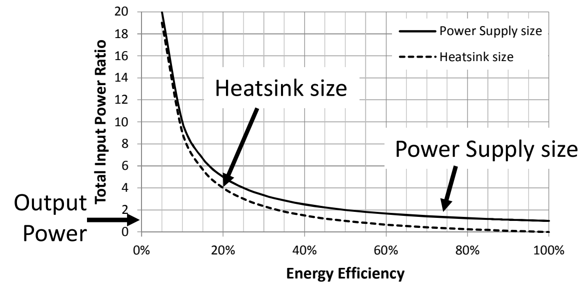


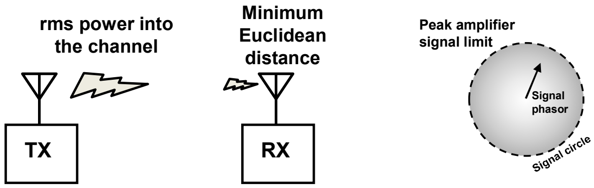
Fig 1. With circuit efficiency below 10%, the cost of power supplies and heatsinks dominates. All power levels in this chart are normalized to the PA output power.

This paper is organized in five sections. Section II looks at the fundamental parameters needed for an efficient and robust link. Section III examines transistor behavior variations which appear at higher operating frequencies. Section IV proposes signal modulation designs that satisfy all of the constraints brought out in the prior sections. Conclusions are drawn in Section V.

II. EFFICIENT COMMUNICATION LINK DESIGN

For a robust digital wireless link, two things are required: 1) the transmitter must supply its available power into the channel, and 2) the minimum Euclidean Distance (MEuD) between symbol waveforms must be as large as possible so that the receiver easily operates effectively. These two link objectives are illustrated in Fig. 2a. Any PA is a peak-power limited circuit, as shown in Fig. 2b, and must be designed to provide the signal's maximum possible peak envelope power (PEP) within the signal circle. Link coverage and range in any communication system analysis are based on signal root-mean-square (rms) power, usually called the average signal power. If the PEP exceeds this average power then a peak-power to average-power ratio (PAPR) exists. Because PEP is fixed by the PA design, any PAPR reduces available transmit power into the channel. This goes directly against link objective #1.

Regarding objective #2, this calls for signal states to be uniformly distributed throughout the signal circle. Common practice, to inscribe a square (QAM) within the signal circle, is clearly suboptimum. Because the use of square QAM is so widespread, this signal type is a useful benchmark to evaluate alternatives against (Section IV).



(a) (b)

Fig 2: Fundamental link design objectives and constraints: a) at the transmitter and receiver individually, and b) peak power limitation at the transmitter is set by the PA design.

III. High Frequency Operation Effects

Transistors cannot operate at infinite speed. A parameter related to the speed that a transistor can operate at is the transition frequency (f_T). When the operating frequency is well below f_T, then the transistor can be considered as having infinite speed and the actual speed restriction can be neglected. The question remains as to when this simplifying assumption no longer is valid.

An initial look into this question, regarding achievable energy efficiency of power amplifiers, is provided in [5]. The result from this simplified analysis is that efficiency losses become measurable when the operating frequency exceeds 5% of f_T. The reason is illustrated in Fig. 3 for an amplifier operating at its peak, saturated output power (P_{SAT}). Transitions between OFF and ON transistor states are limited by f_T. As the operating frequency increases, these finite transition times get closer together. When the transitions begin to overlap then the transistor begins to operate in a nonlinear condition called slew-rate limiting.

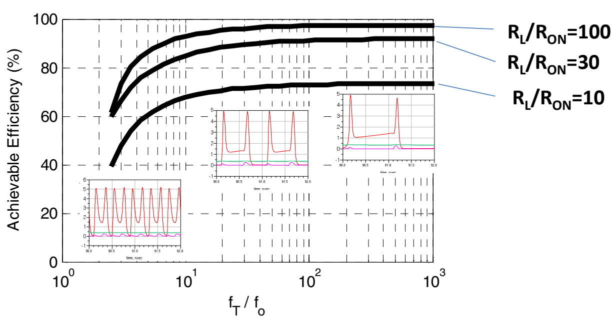
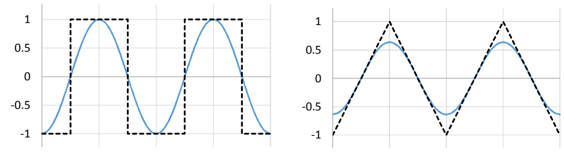


Fig. 3. An amplifier in full power compression exhibits fixed-time transitions between switching states and operates at very high efficiency. At higher operating frequencies these transitions occur closer together, dissipating more power and reducing efficiency.

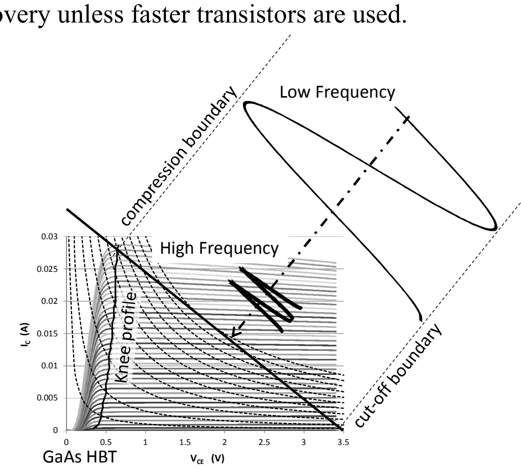
Operating in slew-rate limiting has another effect, seen in Figure 4b, where the maximum amplitude of a sinewave becomes suppressed compared to the maximum available when operating well below f_T (Fig. 4a). Thus a second problem arises: as the operating frequency gets closer to f_T then transistor speed, and not supply voltage, sets the largest available output signal. P_{SAT} drops, and if there is any signal PAPR then average output power drops further, even for the same transistor size. The transition to this suppression of linear signal magnitude theoretically occurs at $f_T /2\pi$, or 16% of f_T.



(a) (b)

Fig. 4. Relating output signal capability and transistor speed: a) transistor has much higher speed capability (dashed curve) than the operating frequency; b) transistor slew-rate limiting begins to suppress the largest available linear signal.

When output signal magnitude is bounded by slew-rate limiting, not only does the output power drop but the PA power dissipation increases because the signal cannot extend to the lower power dissipation regions of the linear amplifier. This is illustrated in Fig. 5. The net result is that PA efficiency falls dramatically, with no possibility of recovery unless faster transistors are used.



(a)

Fig. 5. A large output signal encounters linear output conditions that intersect contours of reduced transistor power dissipation. At higher frequencies the reduced output signal stays in the highest power dissipation region of the load line.

In [6] a transistor is operated at 46% of f_T. Evaluation of the linear operation capability of this amplifier is seen from the slope-gain analysis of this design, repeated here in Fig. 6 from [6, Fig. 8]. Looking at the high bias condition data, the slope gain drops to 0.9 (1.0 is ideal) at -27 dBm input signal power, down 12 dB from saturated operation at -15 dBm input power. 12 dB is a huge amount of output back-off (OBO) to achieve linear operation. Added to the signal PAPR, the actual available output power for communications is P_{SAT} – (OBO + PAPR). If OBO = 12 dB and PAPR = 10 dB, then only 0.6% of P_{SAT} is available to actually use for communications. 99.4% of PA available power is not available to the communication.

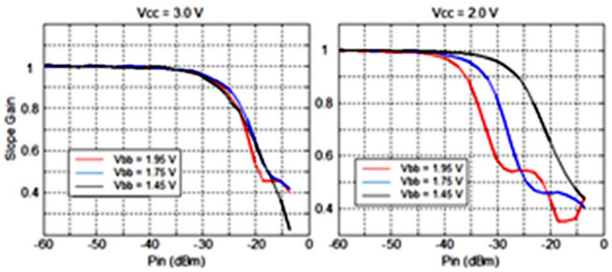


Fig. 6. Measured slope gain across several bias conditions from [6]: linear operation occurs when the slope gain evaluates to 1. Slope gain begins to degrade between 12 and 25 dB below P_{SAT}, here represented by a maximum input power.

Performance limits from transistor f_T are encountered between when the operating frequency is 5% and 10% of f_T. For linear operation at 30 GHz, this means that the implementing transistor technology must have, at full power, f_T values of 300 – 600 GHz or higher. Power transistors today exhibit typical f_T values for various technologies of: LDMOS, 12-20 GHz; GaN, 60-150 GHz; GaAs, 100-300 GHz; SiGe, 200-340 GHz. Only SiGe gets within the lowest part of the desired range. All others operate with various amounts of slew-rate limiting, compromising their linear performance.

IV. Proposed Modulations

A fundamental tenet of any successful standard is that it must be cost effective to implement and produce, providing capability to the system that is far greater than its costs. From a business leader's point of view, one must use more of the available technology that is cheap, and less of the technology that is expensive. This should be obvious. Any system design must allow implementers to follow this path.

In the communication technology business, switching circuits (primarily digital gates) are very cheap to manufacture. The most expensive features to implement are high output power, and tight linearity specifications. Our modulation selections must not require much of these expensive features.

Linearity requirements are largely driven by two waveform features: 1) envelope-zero events, sometimes called zero-crossings, and 2) high order signals where different symbol waveforms are so similar that any distortion of those waveforms causes the receiver to mistake one for another. We desire then to have moderate signal order signal types (to support bandwidth efficiency needs) that preferably have few or no envelope-zero events.

The other expensive feature to implement is power, which means peak power. Given that system coverage is governed by the signal average power, we need the average power to be close to the peak power capability of the PA to meet the robust link objective #1. In other words, PAPR must be small. The smaller the better, so that available PA power gets into the channel. 0dB is best, though a few dB is acceptable.

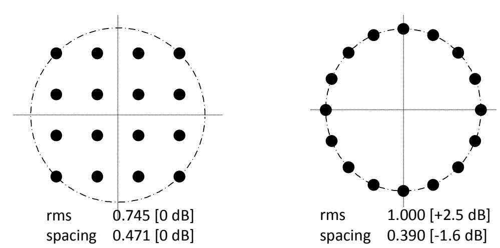
These desires are not at all consistent with OFDM based signal types. The OFDM signal structure intends, by design, to make radio resource management (a digital function) and tolerance of channel delay spread (another digital function) easier [7]. The 'price' is that for N tones, each with M_{SUB} bits per symbol, the signal order M

$$M = 2^{N \cdot M_{SUB}} \tag{1}$$

gets extremely high which requires tight circuit linearity. Further, the PAPR of the summed modulated tones also gets very high, to 10 dB or more. Hardware implementers must use more of what is expensive, and less of what is cheap. And in the presence of slew-rate limiting, both output power and circuit linearity get even more expensive. Any OFDM-type signal is presently not a good choice for mmW operation.

Even the simple QPSK signal has envelope-zero events, which require use of linear amplifiers. Eliminating these removes the need for circuit linearity. This can be done with a change to the signal filtering approach, illustrated here using 16 QAM as a baseline.

Fig. 7a shows conventional square 16 QAM, and conventional 16 PSK is shown in Fig. 7b. From the point of view for link objective #2, the MEuD of 16 QAM is clearly longer so it would be favored. Both constellations, when linearly filtered in quadrature, have envelope-zero events which require linear circuitry. The 16 PSK constellation though has all of its points on the signal circle, so it has higher average power and therefore is more aligned with link objective #1. A quantitative comparison between constellations for overall link performance is found in [8], which proposes a Figure of Merit (FoM) that is the product of the constellation rms value and its MEuD. Larger values of this FoM are more desirable.



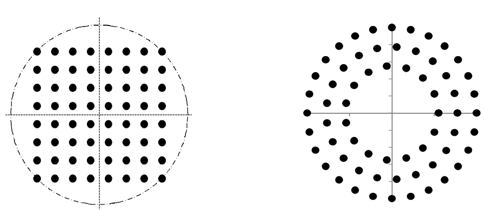
(a) (b)

Fig 7. Comparing two 16-ary signal constellations: a) standard 16QAM; b) 16PSK. Unfiltered constellation rms values, and minimum Euclidean distance values for each are listed.

Compared to 16 QAM, 16 PSK scores lower on MEuD (-1.6 dB) but higher on rms value (+2.5 dB). For overall link operation, the higher rms output power is a greater contribution than the receiver drawback in MEuD. 16PSK is better overall, as long as its constellation rms value is realizable for the actual bandlimited (filtered) waveform. This is readily achieved by changing the filtering from quadrature to polar, which for 16PSK only requires filtering the phase waveform. PAPR then is the ideal value of 0 dB, and all requirement for transmitter circuit linearity is eliminated. For the same PA size, the polar filtered 16 PSK signal provides 2.5 – 1.6 = 0.9 dB better link performance over 16 QAM, all of which is a result of significantly higher power into the channel from the same PA. This is the importance of link objective #1.

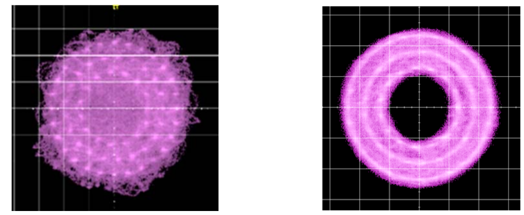
This concept does extend to higher order constellations [9] [10]. For example, consider 64 QAM. Meeting link objective #1 means that the arrangement of the constellation points should be circular instead of square. Spreading the perimeter points of the 64QAM square around the circle spreads them further apart and improves tolerance of noise by 0.9 dB, meeting link objective #2. The remaining inside points are rearranged with equal or greater MEuD into two more inner rings as seen in Fig 8b, forming one of the possible 64-state polar ring (PR) constellations [8]. The rms value of the ring arrangement for these 64 constellation points is 2.2 dB greater than for the square, for the same signal peak value. This becomes "free" power; more power into the channel from the same size PA. The overall link benefit is 2.2 + 0.9 = 3.1 dB, just from rearranging the constellation points. Conventional IQ filtering is certainly possible for PR constellations, and this is shown in Fig. 9a. This filtering yields both envelope-zero events and waveform peaks outside the largest ring. PAPR is nearly 7 dB.

Meeting the objective of eliminating envelope zero crossings is achieved again by changing the signal filtering domain from quadrature to polar. Measured results are seen in Fig. 9b, where the polar-domain Nyquist filtered PR signal (64-PRPF) has no envelope value less than 40% of the peak. All envelope-zero events are removed, and only very small signal peaks remain outside the constellation outer ring. PAPR drops to 2.2 dB. Circuit linearity also goes from being important to becoming not necessary (in line with the linearity-challenged mmW transistors). Economically, the polar filtered ring constellation provides very desirable results.



(a) (b)

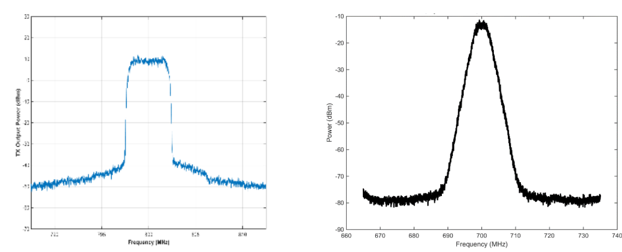
Fig. 8. 64 state constellations (6 bps/Hz): a) standard 64 QAM; b) one of the 64 PR constellations.



(a) (b)

Fig. 9. Measurements of Nyquist filtering the 64PR constellation: a) conventional I and Q linear filtering; b) same filter applied to ρ(t) and ϕ(t) waveforms

Shaping of the signal power spectral density (PSD) is affected by the choice of filtering strategy. Linear filtering using Nyquist impulse responses, which are all sinc() based, does produce largely rectangular PSD in accordance with the Fourier Transform as seen in Fig. 10a. The polar filtering is a nonlinear process, even though the Nyquist impulse response used is identical. This means that the sinc() based filter does not result in sinc() shaped signal voltage waveforms, and therefore the Fourier Transform does not allow the PSD to be of rectangular shape. Rather, the PSD shape is triangular as seen in Fig. 10b. This is very much like the PSD of GMSK, which is also filtered in the phase domain.



(a) (b)

Fig. 10. Spectral measurements of the filtered 64PR constellations of Fig. 8: a) conventional quadrature filtering; b) polar filtering

This wider PSD at lower density values is not due to spectral regrowth, unlike the PSD spreading seen in Fig. 11 showing an overlay of several 5G-NR candidate signals at both the input (left) and output (right) of a linear PA as is typically used in LTE handsets. This spreading is wider than that of the polar filtered signal.

From a millimeter-wave perspective, since most transistors have very restricted linear range and lower than expected output power resulting from slew-rate limiting, having a signal that needs no amplifier circuit linearity and can operate at P_{SAT} is a major benefit to mmW communications applications.

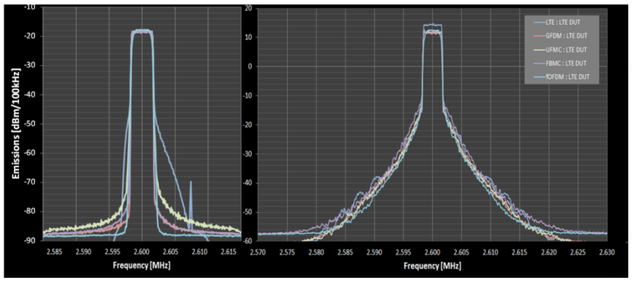


Fig. 11. Spectral regrowth of 5G-NR signals through conventional linear power amplifiers, with the input signals on the left and output signals on the right (from [1]).

A common approach to reducing PSD spreading due to spectral regrowth is to add digital pre-distortion (DPD) around the amplifier. This does work, but it also draws power that does not contribute to a larger output signal. This effectively reduces the efficiency of the PA further in accordance with

$$\eta_{NET} = \frac{P_{OUT}}{\frac{P_{OUT}}{\eta_{PA}} + P_{LIN}} \tag{2}$$

Where η_{NET} is the overall efficiency including the linearizer power P_{LIN}, for a linearized PA providing the output power P_{OUT} at a linearized efficiency of η_{PA}.

Several plots of (2) are provided in Fig. 12, where each curve corresponds to a particular value of P_{OUT} within {0.2, 0.5, 2, 5, and 20 watts}, and the linearized PAefficiency is 37%. This set of curves shows that in order to not degrade the overall efficiency below 35%, the power drawn by the linearizer must not exceed 15-20% of P_{OUT}. Present DPD linearizers draw around 1 watt or more, so for a 0.2 watt mmW amplifier this approach is not viable.

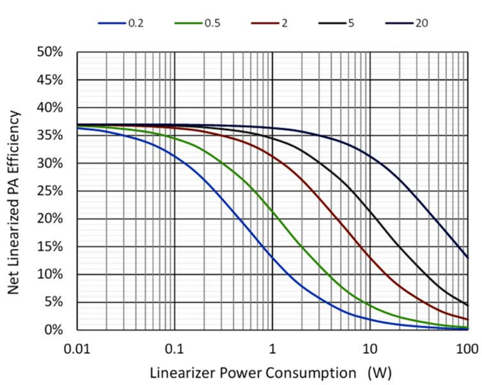


Fig. 12. Impact of linearizer power consumption to the overall efficiency of the linearized PA.

V. Summary

Transmitters operating at millimeter-wave frequencies are not simply higher frequency versions of transmitters we use at microwave frequencies. Once the signal frequency exceeds 5% of the power transistor f_T, additional effects arise that make both circuit linearity and power generation much more difficult. Any need for circuit linearity by the modulation type used forces the output power down further. Unless the mmW output power is 5 watts or higher, it is not practical to apply DPD linearization.

It is possible to design modulation types that are compatible with the operating restrictions of transistors at these high frequencies. The polar filtered 64PR constellation simultaneously achieves bandwidth efficiency of 6 bps/Hz and exceeds 60% energy efficiency. Not needing circuit linearity to achieve this is fundamentally compatible with the lack of circuit linearity from transistors operating at mmW frequencies. And the output power can approach P_{SAT}, without any need for the OBO needed to reach the region of transistor linear performance at mmW. This shows that a lot of communication system performance improvement over presently implemented systems remains available to achieve.

References

[1] D. Brunel, "3GPP 5G Standardization Status RF and mm-Wave Challenges of New Radio," *5G Forum, CICC 2017*, May 1, 2017

[2] H. Wang, et.al. "Towards Energy-Efficient 5G Mm-Wave Links: Exploiting Broadband Mm-Wave Doherty Power Amplifier and Multi-Feed Antenna with Direct On-Antenna Power Combining," *Proc. of BCTM 2017*, pp. 30-37

[3] E. McCune, "Power Amplifier Efficiency Ceilings due to Signal Modulation Type," *Proc. of the 2017 European Microwave Week* (EuMW 2017), Nuremberg, October 2017

[4] E. McCune, D. Babić, R. Booth; D. Kirkpatrick, "Decade Bandwidth Agile GaN Power Amplifier Exceeding 50% Efficiency," *Proc. of MILCOM 2015*, Tampa FL, October 2015, pp. 541-546

[5] E. McCune, "Fundamentals of Switching RF Power Amplifiers," *IEEE Microwave and Wireless Components Letters*, Dec. 2015

[6] Z. Hu, G. Sarris, C. De Martino, M. Spirito and E. McCune, "Design and Linearity Analysis of a D-band Power Amplifier in 0.13 µm SiGe BiCMOS Technology," *Proc. of the 2017 IEEE Compound Semiconductor IC Symposium* (CSICS), Miami FL

[7] E. McCune, "This Emperor has No Clothes?", *IEEE Microwave Magazine*, June 2013, pp. 48-62

[8] E. McCune, "Signal Design and Figure of Merit for Green Communication Links," *Proc. of the 2017 Radio and Wireless Symposium* (RWS2017), Jan 15-17 2017, Phoenix

[9] S. Miller, R. O'Dea, "Peak Power and Bandwidth Efficient Linear Modulation," IEEE Trans. Communications, Vol. 46, No. 12, Dec 1998, pp. 1639-1648

[10] G. Foschini, R. Gitlin, S. Weinstein, "On the Selection of a Two-Dimensional Signal Constellation in the Presence of Phase Jitter and Gaussian Noise," *Bell Sys. Tech. Journal*, Vol. 52, No.6, Jul-Aug 1973