There Is Not Enough Electricity to Run 5G - Finding the Road to 6G

Earl McCune
Professor, Sustainable Wireless Systems, TU Delft, Delft, Netherlands
CTO, Eridan Communications, Mountain View, California
Chair, Energy Efficient Communications Hardware (EECH) Working Group, IEEE Standards

Abstract: The 6G network must be profitable to manufacture, install, and operate. To do this 6G must achieve a more global optimization within the constraints of spectrum, energy, and money than both LTE and 5G-NR. To guide this discussion for improving the radio access network (RAN) with respect to lower energy use and lower implementation cost while maintaining bandwidth efficiency, the metric of link energy efficacy is used. This metric shows that the most significant parameters for achieving high link energy efficacy are: low operating frequencies, clear channel propagation, highly directive antennas, single carrier signals, and low PAPR modulation with no envelope zero crossings. All of these important parameters are not what 5G-NR is doing. This opens several important paths for new research to reach higher performance than LTE and 5G-NR while maintaining the needed network energy efficiency.

I. Introduction

Plans for using the 5G network are wide and varied. Knowing that the radio access network (RAN) is dominantly based on the 4G – Long Term Evolution (LTE) signal type, with which there is now a decade of experience with, it is possible to extend this experience with respect to the LTE networks power draw to predict how much electric power will be needed if the 5G network gets deployed at full scale and operates at its then capacity. This extension presently shows that there is not enough power on our planet now to operate such a network.

All wireless communication networks have three primary constraints that they operate under, with the values of these constraints determined by the application that the wireless communication network is serving. These constraints, presented in Figure 1, are the availability of spectrum within which to operate, the availability of power to operate the network, and the availability of sufficient money to do three essential activities: 1) gain access rights to spectrum, 2) install the network (together comprising capital expenditure, capex) and 3) to operate the network (operating expenditure, opex).

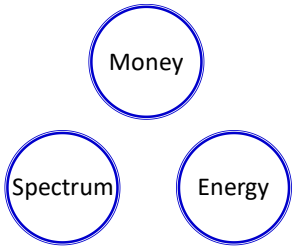

Figure 1. Wireless communication networks operate under three primary constraints: spectrum availability, energy availability, and the availability of money sufficient to both install (capex) and operate (opex) the system.

Presently, the 3GPP standards body has optimized the use of spectrum without major consideration to any required associated requirements on money or energy. As a result it is not presently clear that the 5G network, as specified, will be profitable to build and operate. This leads to the present problem which requires our timely attention, with the objective of achieving a more global solution which allows the coming 6G network to be much more profitable to install and operate.

II. Power Draw

The root of this power draw problem, with regard to the RAN, is shown in Figure 2. On the left is a chart showing the progression of the modulation property of envelope peak-to-average-power ratio (PAPR). The solid curve is a result from linear circuit theory [1] (and its corresponding equation) which limits the available linear amplifier available efficiency depending on the PAPR value of the modulation selected for the associated Standard. The PAPR value has been continuously increasing across the Standard generations. The achieved efficiencies of linear amplifiers for each of these standards in progression is represented by the red ellipses, which are also continuously decreasing but in accordance with the circuit theoretic limit. It is not that development engineering skills are lacking – this is a physical limit.

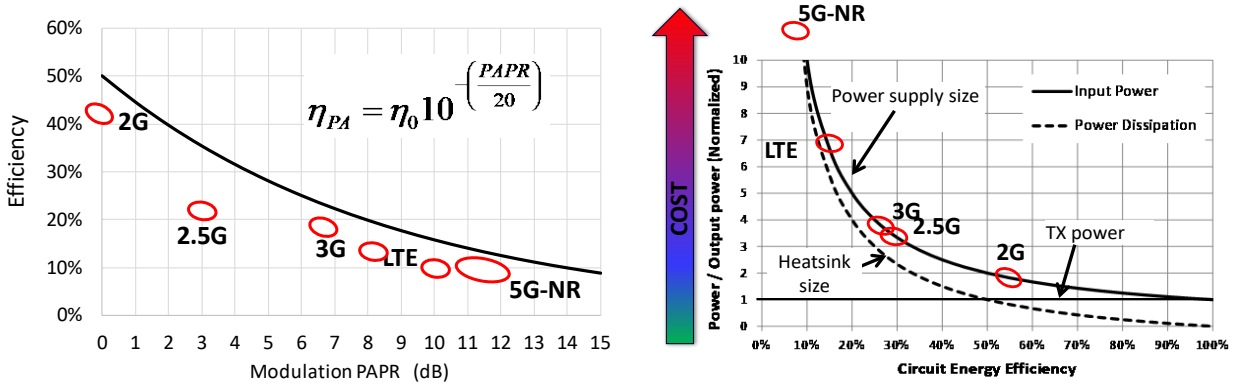

Figure 2. (Left) Signal PAPR increases with each cellular generation correspond to ever-dropping PA efficiency. (Right) Efficiency reductions require corresponding increases in the size (and cost) of power supplies and heatsinks, for identical output power.

Economic connection to this dropping RAN transmitter efficiency is presented on the right chart. For a circuit that is 20% efficient, the manufacturer must provide an associated power supply that is five times larger than the maximum RAN signal output power that the transmitter can provide, and 80% of that power is immediately transferred as heat into the required heatsink for removal from the electronics. This increases capex in the additional cost of both the larger power supply and large heatsink. Opex increases from the additional electricity that must be applied to this power supply. When the circuit efficiency is 10%, the power supply is 10 times larger and 90% of its output is dumped into the heatsink. This characteristic has been tolerated so far, helped that the size of the cellular network was small enough that this cost could be absorbed. Now, the network has grown to the point that its power draw is several percent of total worldwide electricity generation capacity (as measured by carbon footprint) [2] – putting this infrastructure on par with the energy draw of the entire aviation transport industry. At the present 46% compound annual growth rate of the data carried through the cellular network [3], it will draw half of all worldwide

available electricity in 7 more years and all of it in 9 years. Changes are needed, sooner rather than later, to avoid this unsustainable result.

III. Link Energy Efficacy

The need is to be effective in our use of electrical energy (measured here in joules) to move data through the network. Achieving an effect through the use of a different resource is an efficacy, such as the efficacy of lumens per watt used in the lighting industry. Here the interest is the efficacy of bits per joule (1) with additional units-arithmetic provided to expand the efficacy into communication system parameters:

$$\left[\frac{b}{J}\right] = \frac{b/\text{sec}}{J/\text{sec}} = \frac{b/sym \cdot sym/\text{sec}}{watts} \quad . \tag{1}$$

Substituting communication system relationships for each of the units in (1) provides the relationship for link energy efficacy as

$$\chi_{EE} = \frac{\left(N \cdot \log_2(M)\right) f_{sym}}{P_{OUT}/\eta_{PA}} \quad , \tag{2}$$

where N is the number of carriers used, M is the modulation order per carrier, f_{sym} is the modulation symbol rate, P_{OUT} is the average transmit output power, and η_{PA} is the operating efficiency of the transmitter. These parameters are not independent, and among them are all of the design parameters used in wireless link design. When all of the substitutions are complete, we get this relationship between the linear wireless link and its overall energy efficacy [4]

$$\chi_{EE} = \left(\frac{\eta_0 10^{-\left(\frac{PAPR_{dB}}{20}\right)}}{(kTB) \cdot 10^{\frac{NF_{dB}}{10}} \cdot 10^{\frac{SNR_{d,dB}}{10}}} \right) \log_2(M) \cdot N f_{sym} \cdot G_T G_R \left(\frac{\lambda}{4\pi}\right)^2 d^{-p} \quad \left[\frac{b}{J}\right] \tag{3}$$

where NF is the noise figure of the receiver, k is Boltzmann's constant, T is operating temperature in degrees Kelvin, B is the signal channel bandwidth in Hz, G_T and G_R are the linear antenna gains at the transmitter and receiver respectively, SNR_d is the demodulator input signal to noise (power) ratio, λ is the wavelength of the signal carrier frequency in meters, d is the communication distance in meters, and $p \geq 2$ is the environmental effective propagation constant.

For illustration, the uncoded bit-error probability of standard (QAM) modulations for 12 signal order values M are provided in Figure 3 on the left chart. The values used for SNR_d are taken from this chart in the region of the dotted ellipse. The connection between antenna gain values (GT and GR) and their corresponding directivity (i.e. beamwidth) are provided with the chart on the right side of Figure 3.

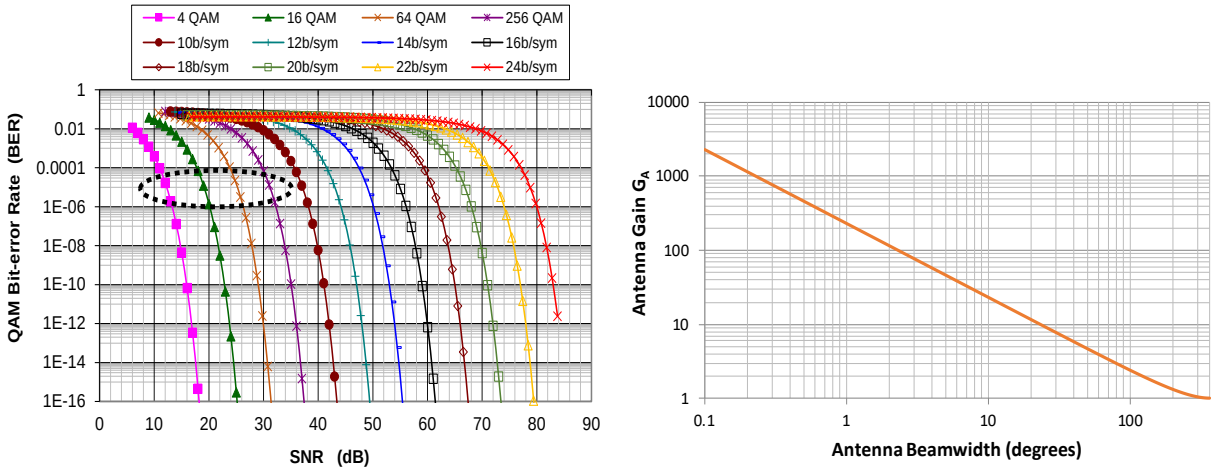

Figure 3. (Left) Bit-error probability vs. SNR (not E_b/N_0) into the demodulator for standard QAM. (Right) Correspondence between antenna gain and beamwidth (directivity) of the antenna.

Contained within the derivation of (3) is the required transmitter power to provide the necessary $SNR_d = 10\log_{10}(snr_d)$ at the receiver demodulator. This power represents the output signal power in the denominator of (2), and is a fraction of the required transmitter peak envelope power (PEP) capability to successfully generate the specified modulation with its PAPR value:

$$P_{out} = \frac{PEP}{PAPR} = \frac{P_R(d)}{G_T G_R \left(\frac{\lambda}{4\pi}\right)^2 \left(\frac{1}{d}\right)^p} = \frac{NF * (kTB) * snr_d}{G_T G_R \left(\frac{\lambda}{4\pi}\right)^2 \left(\frac{1}{d}\right)^p} \quad . \tag{4}$$

IV. Transmitter Physics

There has been a quest across the past century to achieve a linear power amplifier that is also efficient. This quest has never been satisfied, because we now know that physics does not allow it. How this limit comes about is illustrated in Figure 4. On the left is a typical linear amplifier, which must operate along the red bold line (the load line) on this chart. Of particular importance are the dashed curves that represent increasing values of power dissipated in the amplifier transistor as heat (and sent to the heatsink). The signal envelope variations and their corresponding probability density function (PDF) for a 5G new radio (5G-NR) modulation show that the linear amplifier spends most of its time near the middle of the load line, where both the intersecting power dissipation contours have their highest value and the actual available output power is low: both bad for efficiency.

To be efficient, the amplifier must not dissipate much power and so must remain along low value power dissipation contours. This is shown as the green bold curve in the chart on the right side in Figure 4. This only intersects the load line near its endpoints, and is far from the load line where any linear amplifier must operate. This is the dichotomy fought through the past century: a circuit can be linear, or it can be efficient. Choose one. Increasing the efficiency of an amplifier is strictly an exercise in the tolerance of circuit nonlinearity.

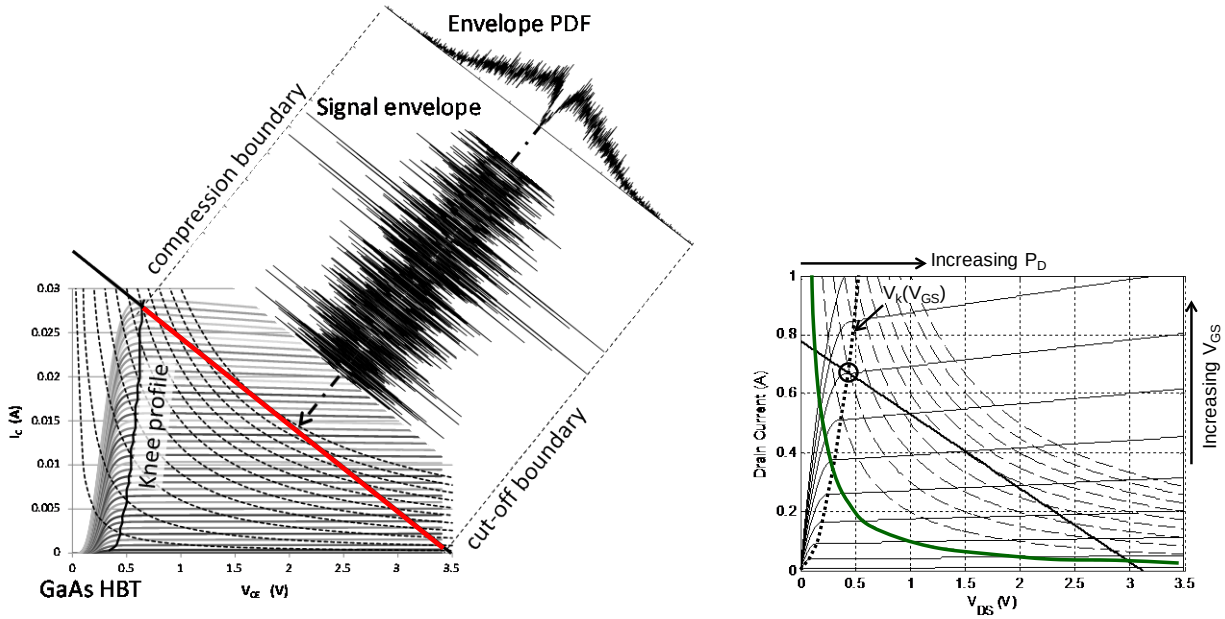

Figure 4. (Left) A linear amplifier must fit the entire signal along the load line. As PAPR increases more time is spent near the middle, where power dissipation is highest. (Right) Efficient amplifiers follow contours of low power dissipation, which are generally far from the load line. Low power dissipation along the load line only occurs at its end-points.

To operate at best efficiency, an amplifier must spend its time at the available load line endpoints and essentially no time along the middle. This is illustrated in Figure 5, primarily by the inset figures with red curves. These red curves peak during the transistor's transition from zero power dissipation (the OFF state) to a minimum power dissipation (the ON state) and back again. This is known as a switching mode of operation. Efficiency is highest when the amount of time of the transition peaks is negligible to the time of the OFF and ON states. As operating frequency increases, these transition times become significant and the available efficiency begins to drop. Eventually the transitions begin to overlap when the operating frequency is high enough. This is the onset of slew-rate limiting, where the amplifier is neither linear nor switching. In this region the available efficiency rapidly decreases.

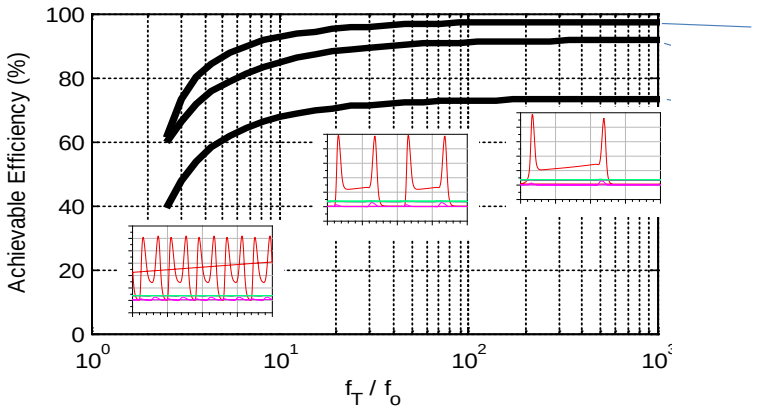

Figure 5. Switching operation rapidly from one end of the load line to the other is effective when the transistor speed is more than 20 times higher than the operating frequency. As operating frequency increases these transitions get closer together and efficiency drops as power dissipation increases.

When the amplifier operates as a switch, it actually ceases to be an amplifier because there no longer is a relationship between variations on the input signal magnitude and any variation in the output signal magnitude – because there is no variation in output signal magnitude. The traditional amplifier model as a two-port circuit, shown on the left in Figure 6, no longer applies. To regain access to output signal magnitude variations it is necessary to provide a third port, such as the supply voltage itself, as the means to produce output magnitude variations

Figure 6. Amplifiers are traditionally treated as two-port circuits with a separate supply voltage (Left). Higher efficiency operation is accessed by treating the supply voltage as a separate independent input (right).

Results of the realizable transmitter efficiency when three-port operation is used are presented in Figure 7. When switching operation is used, the theoretical limit of linear operation no longer applies and significant improvements are measured. For example, LTE improves from 12% to 50%, just over a 4x improvement. But the improvement is actually more than this, because the power supply size went from 7x the output power to 2x the output power, and the all-important dissipated power went from 6x the output power to 1x the output power – a factor of 6 reduction.

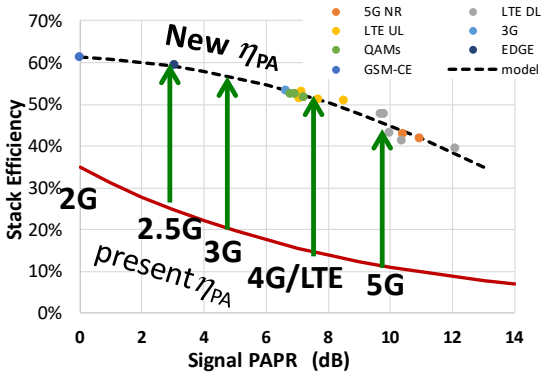

Figure 7. Three-port operation increases efficiency of transmitters, as long as the operating frequency is 5% of f_T or lower.

Increasing the transmitter operating frequency, from FR1 to FR2 in the 3GPP 5G Standard, incurs additional effects. Figure 8 shows the impact on transistor available gain as its operating frequency gets closer to its transition frequency f_T. More than 25 dB is lost when transitioning from 1.5 GHz (in the FR1 microwave band) to 28 GHz (in the FR2 millimeter-wave band). This gain must be recovered since signal levels are specified and fixed, requiring more circuit stages and their operating power be added to the transmitter.

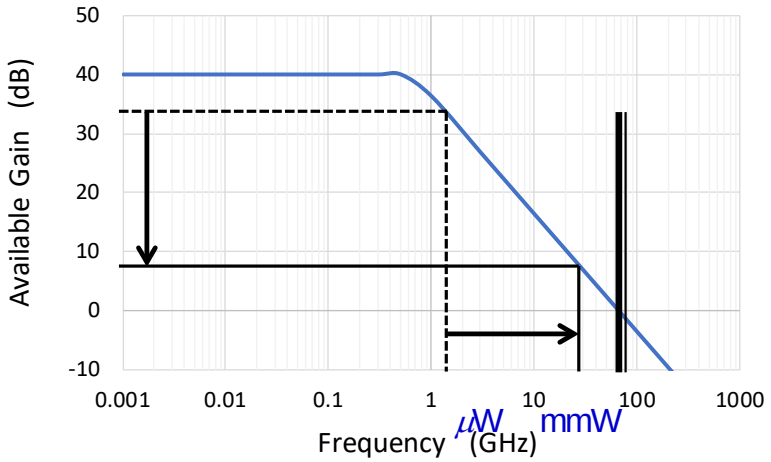

Figure 8. As operating frequency approaches f_T, available gain drops rapidly from any transistor.

Further, as seen in Figure 5 the transistor at higher operating frequencies enters slew-rate limiting. This adds the consequences seen in Figure 9. The left side chart shows the behavior when operating at frequencies well below f_T and the transistor does switch. Bur when the transistor operates in slew-rate limiting, the chart on the right in Figure 9 shows that the output signal peak power drops not only to the slew-rate, but also drops more when the operating frequency is further increased. Low available gain and low available output power combine into very inefficient transmitter implementations.

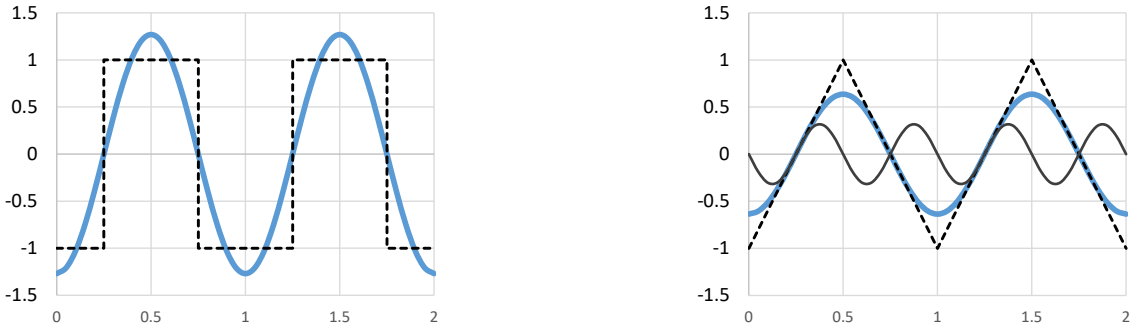

Figure 9. At low enough frequencies where the transistor is a good switch, transition times are negligible and output power is high (left). When the operating frequency is high enough that slew-rate limiting begins, available output power is reduced and drops further with continued frequency increases.

V. Modulation Selection Effects

The selection of signal modulation for any Standard has a significant impact on the achieved link energy efficacy [5]. The important parameters are whether the modulation is single-carrier or multi-carrier, whether there are envelope zero crossings, and what the PAPR of the modulation is. For example, if the signal is multi-carrier, any circuit nonlinearity causes cross-modulation among the carriers which is unacceptable, so linear circuitry is required. LTE and 5G-NR are multi-carrier signals, so we are justified in evaluating linear amplifier performance with modulated signals here.

With the availability of switching transmitters at lower frequencies, within FR1 these transmitters can be built based on Nyquist Sampling Theory instead of linear network theory. We focus here on FR2 where the likelihood of transistors operating in slew-rate limiting is high. For an example we take the 60 GHz amplifier described in [6]. Data from measurements of this amplifier are presented in Figure 10. On the left side, the published data on the top chart is evaluated for its linear operating region at the bottom which occurs when there is 1 dB of output power change when the input power change is 1 dB. This is seen to occur when the amplifier input power is below -12 dBm. PEP is defined at the input power where the amplifier distortion is at the limit allowed for the signal. The distortion of the amplifier is (1 – slope power). Taking this limit at 3%, the PEP is set at -10 dBm input power, equaling +12 dBm output power. This is an output back-off (OBO) of (17 – 12) = 5 dB, corresponding to 11 dB input back-off. This is well below the P1dB point, which has a 20% distortion. The corresponding efficiency at this output power is 5%.

The right-hand charts in Figure 10 show the impact of the 5G-NR CP-OFDM modulation on the overall efficiency of this amplifier. The 5% efficiency at PEP is η_0 for the limit curve in Figure 2 (left), which at 10 dB PAPR here shows the operating efficiency is 1.6%.

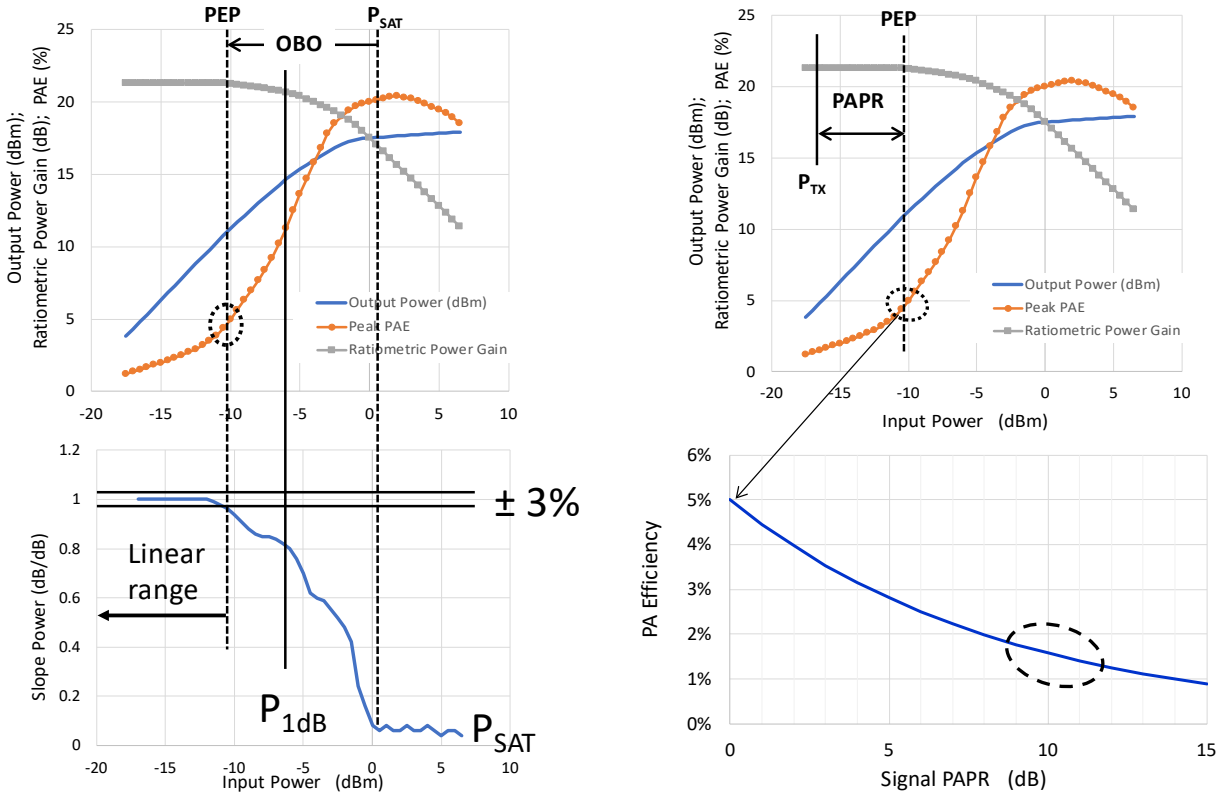

Figure 10. Output power back-off (OBO) is necessary to access sufficiently linear operation from an amplifier (left). This sets the maximum peak envelope power (PEP) that a modulated signal can have while maintaining sufficient linearity.

VI. 6G Options

To explore the road to 6G, we begin by generating the plot in Figure 11 of results from (3) for LTE links across distance for p = 3.2 and at three carrier frequencies: 750 MHz, 1900 MHz, and 28 GHz. The results are shown by the solid lines, and the link efficacy varies by 10 orders of magnitude across 10 m to 10 km link distance. There is also a nearly 4 order of magnitude drop in efficacy across these frequencies, all else being equal. Also included on this chart are the corresponding transmitter output powers from (4), which grow at the same rate the link energy efficacy falls, confirming the importance of using actual transmitter power in any evaluation of link energy efficacy.

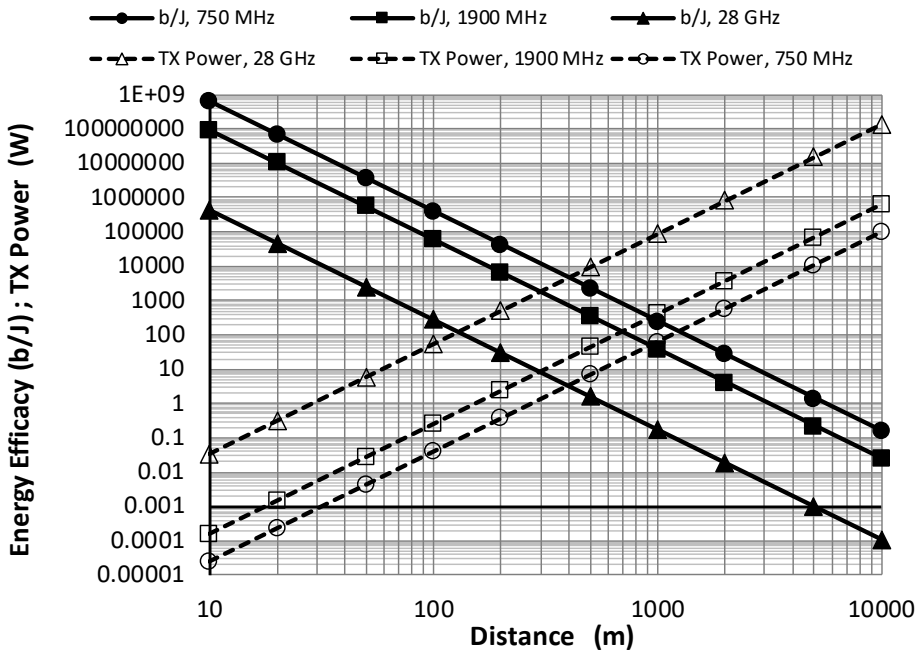

Figure 11. Link Energy Efficacy values (solid lines) for various frequencies of present-day LTE operation across distance in a general suburban environment. Corresponding transmitter powers needed to operate the link (dashed lines) account for the drops in link energy efficacy.

Working on the impact of signal modulation selection, we propose a modulation selection change for the FR2 bands as shown in Figure 12. Instead of CP-OFDM based on 16-QAM subcarriers, which is a present 5G-NR case with a maximum bandwidth efficiency of 4 bps/Hz, we propose a pure PSK (pPSK) with 16 states that also maximizes at the same bandwidth efficiency. All pPSK transitions also remain on the constellation circle, so PAPR = 0 dB. There is no need for amplifier linearity, so this amplifier can now be operated directly at its saturated output power (P_{SAT}). Output power into the channel increases from 5 dBm to 17 dBm, and the operating efficiency increases from 1.6% to 20%. All with no change in radio hardware. Just a change in the adopted modulation.

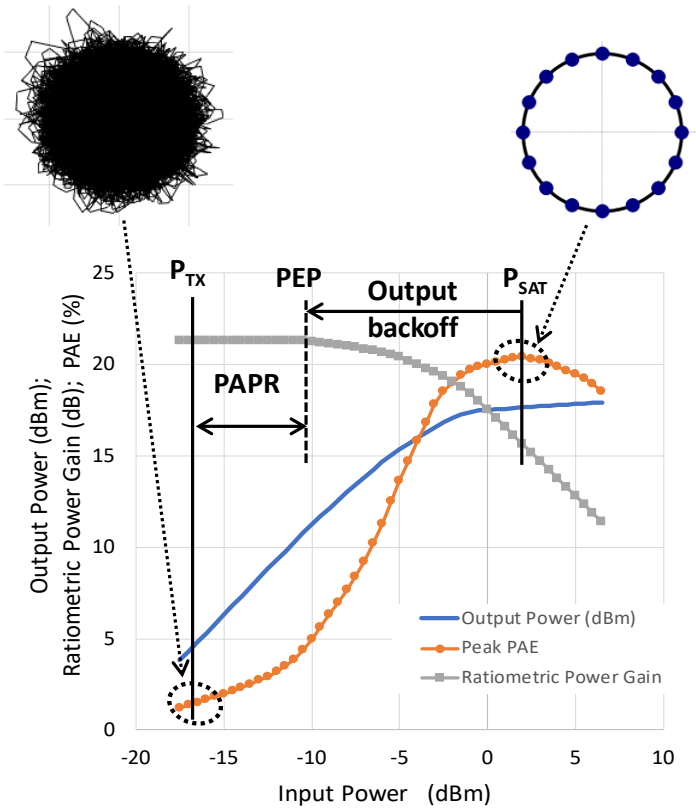

Figure 12. Changing link modulation from CP-OFDM (presently the 5G-NR standard) to a constant envelope PSK immediately increases both output power (here by 12 dB) and efficiency (here by 10x, ~2% to 20%) while maintaining bandwidth efficiency.

The economic consequence of this modulation change is presented in Figure 13. The power supply for 5G-NR is 62.5 times larger than the output power from the transmitter. With this modulation change the overscaling factor drops to 5. Heat into the heatsink drops from 61.5x the output power to 4x the output power. This represents major cost reductions in both hardware manufacture and in operating costs. Research in improving FR2 amplifier P_{SAT} without need for corresponding need for circuit linearity will directly improve this result.

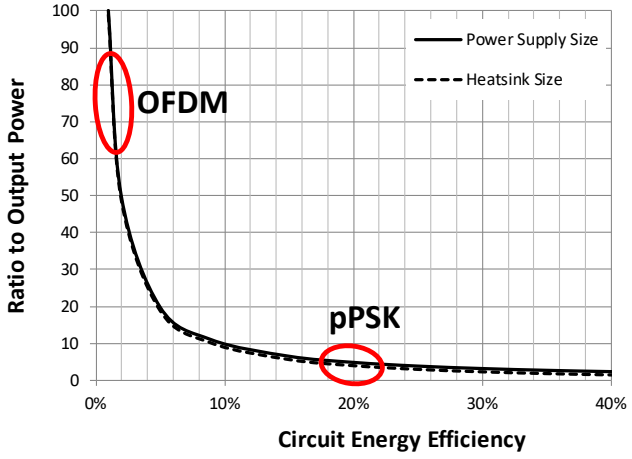

Figure 13. Economic results of the modulation change show that both the power supply and heatsink shrink from more than 80x oversized to 5x oversize.

These improvements do not yet get to an improvement of 10000x over the LTE baseline link energy efficacy. Exploration of ways to do this are presented in Figure 14.

This chart is based on the present Link energy efficiency for LTE at 1900 MHz with propagation p = 3.2. Based on this case performance, an improvement of 10000x is presented as a line labeled '6G goal'. The proposed 1900 MHz cases are

- Present LTE baseline (solid squares)
- Adopt DSR technology (solid circles)
- Use high antenna directivity at both ends of the link (solid triangles)
- Change to a near-zero PAPR modulation (open circles)
- Arrange the network to have less environment loss (open triangles)
- 28 GHz FWA with directivity at both ends and less environment loss (open squares)

Largest improvements come from two-sided antenna directivity and from changing to pPSK modulation. Yet, there is no solution apparent that operates beyond a 200 m communication distance. The strong impact of directional antennas is well aligned with SDMA as used in massive MIMO. The short-range installation is consistent with the installation of small cells. Though this chart shows that the entire network will need to be implemented with small cells to achieve this level of link energy efficacy.

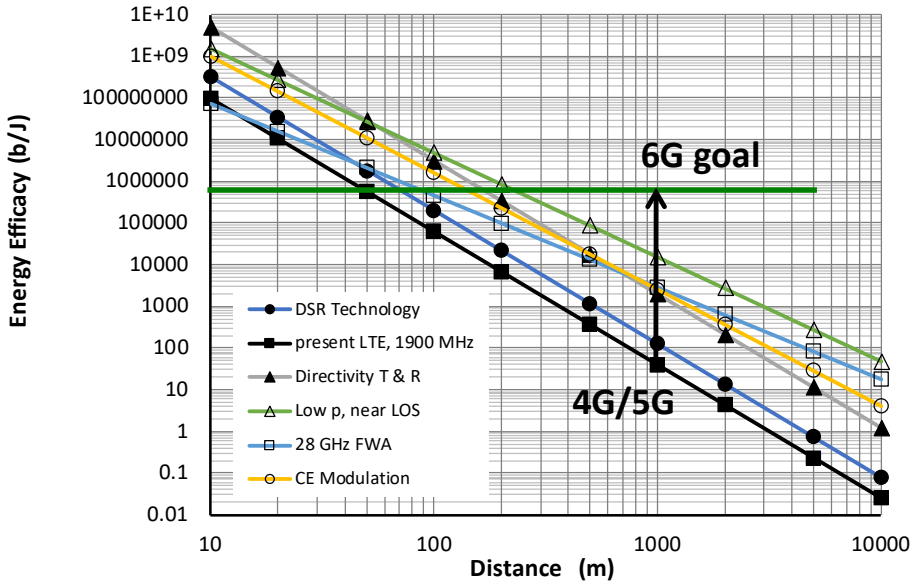

Figure 14. Improving link energy efficacy by a factor of 10000 above present-day values involves a combination of 1) adopting DSR techniques for hardware implementation, 2) directional antennas (better if at both ends), 3) reducing communication distance, and 4) changing modulation to constant envelope while maintaining bandwidth efficiency.

VII. Conclusions

6G must achieve a more global optimization within the constraints of spectrum, energy, and money. The 6G network must be profitable to manufacture, install, and operate. To guide the discussion for improving the RAN with respect to lower energy use and lower implementation

cost while maintaining bandwidth efficiency, the metric of link energy efficacy is used. This metric shows that the most significant parameters for achieving high link energy efficacy are

- Low operating frequency
- Clear channel propagation
- Highly directive antennas
- Single carrier signals
- Low PAPR modulation with no envelope zero crossings

It is a surprise that adopting higher order modulations does not rank high, though the link energy efficacy metric (3) shows that the increase in signal bandwidth efficiency comes along with a need for higher SNR into the receiver demodulator, which requires higher transmitter power and a larger power draw that largely compensates the bandwidth efficiency increase.

References

[1] S. Miller, R. O'Dea, "Peak Power and Bandwidth Efficient Linear Modulation," *IEEE Trans. On Communications*, Vol. 46, No. 12, Dec. 1998

[2] T. Klein, "GreenTouch Consortium: Transforming ICT Networks for a Sustainable Future," *GreenTouch Overview Deck January 2013.pdf*, presentation available at https://www.yumpu.com/en/document/read/37522081/greentouch-overview-deck-january-2013pdf, accessed on 8 March 2020

[3] Cisco Systems, available at https://www.cisco.com/c/m/en_us/solutions/service-provider/forecast-highlights-mobile.html, accessed 8 March 2020

[4] E. McCune, "A Comprehensive View of Wireless Link Energy Efficiency", *Proc. of 2019 IEEE Global Communications Conf.* (GLOBECOM), 9-13 Dec. 2019, Waikoloa, HI

[5] S. R. Biyabani, R. Khan, M. M. Alam, A. A. Biyabani, E. McCune, "Energy Efficiency Evaluation of Linear Transmitters for 5G NR Wireless Waveforms," *IEEE Trans. On Green Communication and Networking*, vol. 3, Issue: 2, June 2019, pp. 446-454

[6] M. Babaie, R. B. Staszewski, L. Galatro, M. Spirito, "A wideband 60 GHz class-E/F2 power amplifier in 40nm CMOS," *Proc. of the 2015 IEEE Radio Frequency Integrated Circuits Symposium* (RFIC), paper RMO4B-4, pp. 215-218