Modulation Scaling for Energy Aware Communication Systems

Curt Schurgers
NESL, EE Dept., UCLA[†]
curts@ee.ucla.edu

Olivier Aberthorne
NESL, EE Dept., UCLA[†]
thornado007@yahoo.com

Mani B. Srivastava
NESL, EE Dept., UCLA[†]
mbs@ee.ucla.edu

ABSTRACT

In systems that require low energy consumption, voltage scaling is an invaluable circuit technique. It also offers energy awareness, trading off energy and performance. In wireless handheld devices, the communication portion of the system is a major power hog. We introduce a new technique, called modulation scaling, which exhibits benefits similar to those of voltage scaling. It allows us to trade off energy against transmission delay and as such introduces the notion of energy awareness in communications. Throughout our discussion, we emphasize the analogy with voltage scaling. As an example application, we present an energy aware wireless packet scheduling system.

Keywords

energy awareness, adaptive modulation, scaling

1. INTRODUCTION

In tetherless battery-operated devices, power consumption is a critical design aspect. It has been realized that it is **energy awareness**, in addition to low power, that is required for most applications [1]. Scaling the supply voltage is the most common circuit technique to offer both low energy consumption and energy awareness [2]. In operating system research, the clock speed and supply voltage are dynamically adjusted based on the predicted workload [3]. Another approach, proposed for self-timed [4] and synchronous [5] systems, is to use the amount of buffered load to steer the adaptation.

Furthermore, a lot of these battery-operated devices are equipped with a wireless communication subsystem. A major source of their energy consumption is the actual data transmission over the air. Despite the work on energy awareness in digital electronic circuits, it has been overlooked that the same tradeoffs are present in communications as well. In this paper, we show that the **modulation can be scaled** much the same way as operating voltage can, reducing the overall energy consumption for transmitting each bit. Although the basic idea of changing the modulation on the fly has been used to increase the throughput in the presence of fading channels [6], it has never been exploited for low power purposes. We have applied this principle towards an **energy aware wireless scheduling system**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ISLPED'01, August 6-7, 2001, Huntington Beach, California, USA.
Copyright 2001 ACM 1-58113-371-5/01/0008…$5.00.

2. COMMUNICATION THEORY

Since we investigate the relationship between modulation and transmission speed, we first need to derive the relevant expressions. We focus on Quadrature Ampitude Modulation (QAM) due to its ease of implementation and analysis [6]. However, our techniques are perfectly extendable to other modulation schemes, only the formulas and curves will change accordingly. The performance of QAM in terms of Bit Error Rate (BER) is given by (1)-(3) [7].

$$BER = \frac{4}{b} \cdot \left(1 - \frac{1}{2^{b/2}}\right) \cdot Q\left(\sqrt{3 \cdot \frac{SNR}{2^b - 1}}\right) \tag{1}$$

$$SNR = \frac{P_S}{P_n} \cdot A \tag{2}$$

$$P_n = N_0 \cdot \beta \cdot R_S \tag{3}$$

The constellation size in number of bits per symbol is represented by b. The received Signal to Noise Ratio (*SNR)* is defined as (2), where P_S is the transmit power and A contains all transmission loss components. The noise power P_n is a function of the symbol rate R_s, the noise power spectral density N_0 and a factor β that takes into account all other elements, such as filter non-idealities. [7]. We can manipulate these equations to obtain the following expression for the required transmit power:

$$P_S = C_S \cdot R_S \cdot \left(2^b - 1\right) \tag{4}$$

$$C_S = \frac{N_0 \cdot \beta}{A} \cdot \Gamma \tag{5}$$

$$\Gamma = \left(\tfrac{1}{3}\right) \cdot \left[Q^{-1}\left(\left(\tfrac{1}{4}\right) \cdot \left(1 - \frac{1}{2^{b/2}}\right)^{-1} \cdot b \cdot BER\right)\right]^2 \tag{6}$$

Since our goal is to investigate the energy-delay characteristics while varying the communication parameters we want to keep the system performance constant for a fair comparison. In practical scenarios it makes sense to operate at a target BER. Due to the inverse $Q(.)$ function in (6), C_s is only a weak function of b.

An energy aware communication system **adjusts *b* and *R_s*** to reduce the overall energy. The transmit power P_S (delivered mainly by the power amplifier), however, is not the only source of

[†] Networked and Embedded Systems Lab (NESL), EE Dept., University of California at Los Angeles, 56-125 B, Eng. IV Bldg., UCLA-EE Dept., Box 951594, Los Angeles, CA 90095-1594

power spending. Electronic circuitry for filtering, modulating, upconverting, etc. contributes as well. Equation (7) expresses this component P_E for a system that can dynamically change the symbol rate [8]. Parts of the circuitry operate at a frequency that follows the instantaneous symbol rate, while other parts have a fixed frequency proportional to the maximum symbol rate. The proportionality factors and switching activity are all incorporated in C_A and C_B.

$$P_E = \left[C_E + C_R \cdot \frac{R_{S_{\max}}}{R_S} \right] \cdot R_S \tag{7}$$

$$C_E = C_A \cdot V^2 \qquad\qquad C_R = C_B \cdot V^2 \tag{8}$$

The total power consumption is the sum of both the transmit and electronics power. As in digital circuit design, it makes more sense to look at the energy consumption rather than the total power. We can express the energy to transmit one bit, E_{bit}, as:

$$E_{bit} = (P_S + P_E) \cdot T_{bit} \tag{9}$$

In this equation, T_{bit} is the time it takes to transmit one bit. The goal is to minimize the energy per bit by choosing the correct values of b and R_s. For typical applications, however, we need to constrain the total delay a packet may incur, translating to a bound on T_{bit}. The optimization problem can be summarized as:

$$\min \quad E_{bit} = \left[C_S \cdot \left(2^b - 1\right) + C_E + C_R \cdot \frac{R_{S_{\max}}}{R_S} \right] \cdot \frac{1}{b} \tag{10}$$

$$T_{bit} = \frac{1}{b \cdot R_S} \leq T_{\max} \tag{11}$$

3. PERFORMANCE TRADEOFFS

Our numerical results in this section are based on table 1. The values of C_S, C_E and C_R are extracted from [8], which describes the actual implementation of an adaptive QAM system. Figure 1 depicts E_{bit} as a function of b and R_s as obtained from (10). The corresponding values of T_{bit} from (11) are shown in figure 2. Based on these two figures, we can evaluate the performance in terms of energy consumption for varying constraints on the delay (i.e. varying T_{max}).

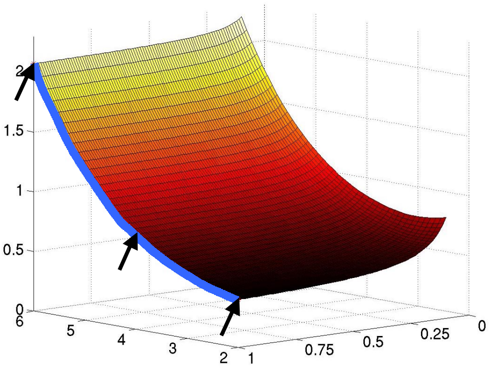

Figure 1: Energy consumption for adaptive R_s system

Table 1: Simulation settings

R_{Smax}	1 MHz
BER	10^{-5}

C_S (b=4)	10^{-7}
C_E	$8.\ 10^{-8}$
C_R	10^{-7}

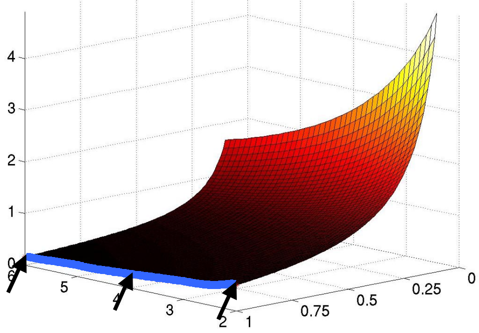

Figure 2: Delay per bit

From these figures, it is clear that operating at the maximum R_S is preferable for any b. This is logical as this results in both a lower T_{bit} and a lower E_{bit}. The symbol rate should therefore be chosen as high as possible, considering implementation issues and their power penalties. Varying the constellation size b is the only option to trade off energy versus delay. In practice, b does not have an infinitesimal granularity but typically only takes on even integers, indicated by the black arrows in figures 1 and 2.

Note that the results of figure 1 are for a communication system that has provisions to vary the symbol rate on the fly. In (7), this introduces the term with constant C_R. Since the optimal symbol rate is always the maximum one, a variable symbol rate provision is not needed for energy awareness reasons. In fact, the **system can be designed for a fixed symbol rate** instead. The circuitry that is described by the term with constant C_R is still present of course. We therefore cannot simply remove this term. However, we modify equation (10) by setting R_{Smax} equal to R_S, such that the energy per bit is now expressed as:

$$\min \quad E_{bit} = \left[C_S \cdot \left(2^b - 1\right) + C_E + C_R \right] \cdot \frac{1}{b} \tag{12}$$

Upon investigating (12), it is clear that E_{bit} is no longer a function of the symbol rate. Since a higher R_s still results in a lower T_{bit}, it is still beneficial to operate at the highest symbol rate that can be implemented efficiently. The reason is that besides the advantage of lower delays, this would also improve the capacity if the wireless medium were shared. We can visualize the energy and delay curves by taking the intersection of the surface in figures 1 and 2 with a plane at $R_S = 1$ MHz.

It is clear that energy and delay can be traded off against each other by varying b. In analogy with voltage scaling techniques in digital circuits, we refer to this process as **modulation scaling**. Depending on the delay that is acceptable, the constellation size can be adapted to meet that constraint with the minimum amount of energy. If this adaptation is performed on the fly, it results in **energy awareness**.

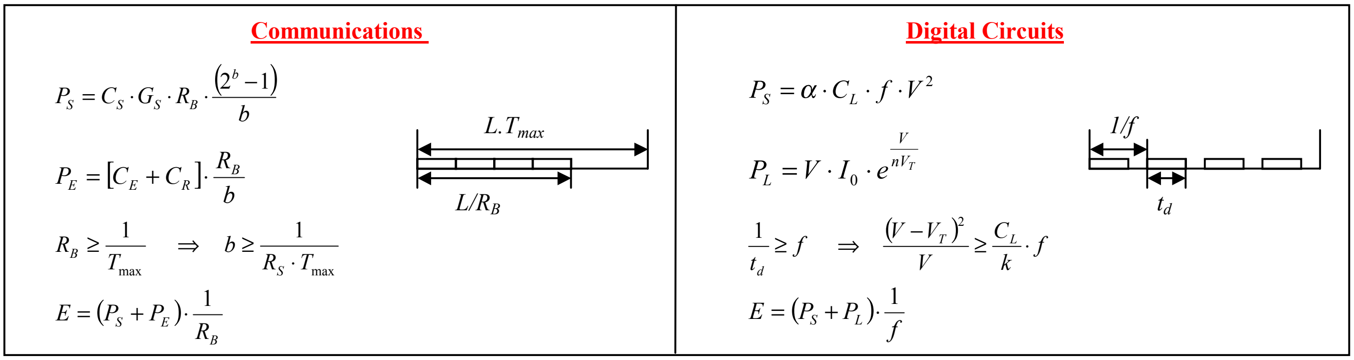
Communications
$P_S = C_S \cdot G_S \cdot R_B \cdot \frac{(2^b - 1)}{b}$
$P_E = [C_E + C_R] \cdot \frac{R_B}{b}$
$R_B \geq \frac{1}{T_{\max}} \quad \Rightarrow \quad b \geq \frac{1}{R_S \cdot T_{\max}}$
$E = (P_S + P_E) \cdot \frac{1}{R_B}$
$L.T_{max}$
L/R_B
Digital Circuits
$P_S = \alpha \cdot C_L \cdot f \cdot V^2$
$P_L = V \cdot I_0 \cdot e^{\frac{V}{nV_T}}$
$\frac{1}{t_d} \geq f \quad \Rightarrow \quad \frac{(V - V_T)^2}{V} \geq \frac{C_L}{k} \cdot f$
$E = (P_S + P_L) \cdot \frac{1}{f}$
$1/f$
t_d

Figure 3: Comparison between adaptive modulation and voltage scaling

4. COMPARISON BETWEEN VOLTAGE SCALING AND MODULATION SCALING

The equations in the previous sections resemble those of voltage scaling, yet there are some key differences. It is important to highlight these differences, as they also contribute to a physical understanding of the tradeoffs of modulation scaling. Figure 3 places both scaling techniques next to each other. In the equations for voltage scaling, P_S is the switching power and P_L the leakage power [3]. It is clear that **the functionality of supply voltage *V* corresponds to that of the constellation size *b*** (hence the terms voltage and modulation scaling). In the left column, the energy is only dependent on b and not on R_B. Equivalently in the right column, the energy term due to the switching power (P_S/f) depends on V and not on f. There is however a crucial difference, regarding the interpretation of time.

In the digital circuit case, the total effective delay for an operation t_d has to be smaller than $1/f$. Similarly in a communication system the total time it takes to transmit a packet (or a bit) has to be smaller than a certain maximum value. **The difference between both systems, however, is the period over which energy is consumed**. In a communication system, the power has to be multiplied by the **effective time** of the operation. In digital circuits, on the other hand, the power is multiplied by the cycle time, which is in effect the maximum delay. As such, there is no true one-to-one mapping between R_B (or R_S for that matter) and f. However, when considering R_B and f as constants of the system, they result in a lower bound on b or V in similar ways (see the third line of equations in figure 3).

5. ENERGY AWARE WIRELESS PACKET SCHEDULING

Like energy aware OS scheduling, we can perform energy aware packet scheduling. We study the communication system setup depicted in figure 4, which consists of a point-to-point transmission link. Packets arrive at the sender and possibly need to be buffered before transmission. We assume that both the packet sizes and the intervals between packet arrivals, called inter-arrival times, follow an exponential distribution. Without modulation scaling, this setup corresponds to the well-known M/M/1 queuing system [9].

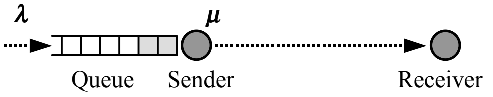

Figure 4: Setup of the queuing system

The average packet arrival rate is denoted by λ. The inverse of the average service time is called the service rate, μ, which gives the average number of packets that can be sent per unit time. It is expressed by (13), where L is the average packet size.

$$\mu = \frac{R_B}{L} = \frac{b \cdot R_S}{L} \tag{13}$$

Because of the statistical properties of inter-arrival and service times, the number of packets in the buffer may vary considerably. Most of the time, the buffer is empty. In those situations, it is beneficial to scale the modulation down to conserve energy. When the buffer starts to fill up, we can increase b to avoid long queuing times or buffer overflow. This kind of system therefore is a good candidate for modulation scaling. A similar observation has been made for digital circuits, where a queue is introduced to average the rate over several samples in a DSP system [5].

The idea is to choose the constellation size based on the number of packets in the system (i.e. being transmitted or in the queue), which we refer to as the system state. For each state $S_n = n$, we have a particular constellation size b_n, which translates into a value of μ_n through equation (13). The collection of $\{b_n\}$ for all the possible states determines the average energy consumption and delay of the queuing system. Our goal is to find which $\{b_n\}$ minimizes the energy for a particular delay constraint. We can analyze this problem using queuing theory. In steady state, the probability of being in state n can be expressed as [9]:

$$P_n = P_0 \cdot \frac{\lambda^n}{\prod_{k=1}^{n} \mu_k} \tag{14}$$

In this equation, P_0 is a constant such that the sum of P_n over all states is equal to 1. We assume an infinite buffer size, which is a reasonable approximation for real systems, as memory has become rather inexpensive for these applications. For each state the energy consumption per bit is given by (12).

The average energy per packet, E_{av}, is the ratio of the average power per packet and the packet arrival rate. The average power is the product of the probability P_n of being in a state, the rate μ_n in that state and the average energy per packet ($E_n.L$) in that state:

$$E_{av} = \frac{P_{av}}{\lambda} = \frac{1}{\lambda} \cdot \sum_{n=1}^{\infty} P_n \cdot \mu_n \cdot E_n \cdot L \tag{15}$$

Since (13) holds in every state, we can simplify this expression to:

$$E_{av} = \frac{R_S}{\lambda} \cdot \sum_{n=1}^{\infty} P_n \cdot E_n \cdot b_n \tag{16}$$

Furthermore, queuing theory tells us that we can express the average delay of a packet as [9]:

$$T_{av} = \frac{1}{\lambda} \cdot \sum_{n=0}^{\infty} n \cdot P_n \qquad (17)$$

For our numerical evaluation and subsequent simulations, we have again chosen the settings of table 1, augmented with those of table 2. In figure 5 each point on the energy versus delay curve represents the average performance of the queuing system for a particular set of $\{b_n\}$. We have only plotted those operating points that minimize the energy for a delay constraint. The dashed curve is for the ideal system that would allow fractional values of *b*. In practical situations, we select *b* from the set of even integers, which results in the curve labeled 'Queuing'. Table 3 gives the values of $\{b_n\}$ for the operating point indicated by the arrow.

However, even this system is difficult to implement in practice, because the modulation would have to be adjusted every time the system state changes. This means that the constellation size can change when either a packet enters or leaves the queuing system. On the other hand, the receiver needs to know what modulation scheme the sender is using in order to decode the symbols and every change in constellation size needs to be communicated to the receiver.

In practice, it is more appropriate to **adapt the constellation size only at the beginning of the packet transmission**. An indicator (encoded with a fixed modulation) in the packet header that describes the modulation used for the packet payload. The modulation scaling is performed based on the number of packets in the queue at the time the transmission starts. The curve labeled 'Simulation' in figure 5 presents the performance of such a practical scheme (it includes the overhead due to the indicator). There is a penalty compared to the theoretical queuing system since the modulation is only adapted when a packet starts being transmitted, instead of every time the number of packets in the system changes. For this practical system, we can select the best operating point for each delay constraint from the curve in figure 5. This operating point defines the values of $\{b_n\}$ that have to be chosen.

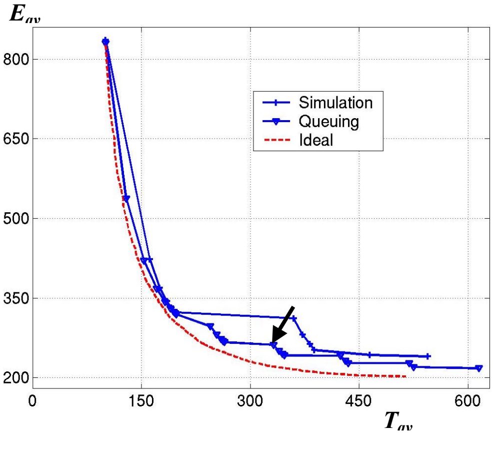

Figure 5: Energy-delay tradeoff for an energy aware queuing system

Table 2: Simulation settings

λ (packets/s)	5000
L (bits)	400
μ_n (packets/s)	$2500 \cdot b_n$

Table 3: Settings for an example operating point

S_n	1	2	3	4	5	≥ 6
b_n	2	4	4	4	6	6

6. CONCLUSIONS

We have presented modulation scaling, which allows us to design energy aware communication systems. We have highlighted the similarities and differences compared to voltage scaling used in digital circuits. A lot of approaches that have been explored in the context of voltage scaling can be applied to modulation scaling as well. We have investigated this for energy aware wireless packet scheduling. However, many other applications can be envisioned that benefit from modulation scaling. Also techniques that improve the system's energy performance can be incorporated into this framework, such as parallelism.

7. ACKNOWLEDGMENTS

This research was supported in part by DARPA under the PAC/C program.

8. REFERENCES

[1] Graybill, R., "DARPA Power Aware Computing/ Communication," *http://www.darpa.mil/ito/research/pacc/.*

[2] Chandrakasan, A., Sheng, S., Brodersen, R., "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, Vol.27, pp. 473-484, Dec..

[3] Govil, K., Chan, E., Wasserman, H., "Comparing Algorithms for Dynamic Speed-Setting of a Low-Power CPU," *MobiCom'95*, Berkeley, CA, pp. 13-25, Nov. 1995.

[4] Nielsen, L., Niessen, C., Sparsø, J., van Berkel, K., "Low Power Operation Using Self-Timed Circuits and Adaptive Scaling of the Supply Voltage," *Trans. on VLSI Systems*, Vol.2, No.4, pp. 391-397, Dec. 1994.

[5] Gutnik, V., Chandrakasan, A., "Embedded Power Supply for Low-Power DSP," *Trans on VLSI Systems*, Vol.5, No.4, pp. 425-435, Dec. 1997.

[6] Ue, T., Sampei, S., Morinaga, N., Hamaguchi, K., "Symbol Rate and Modulation Level-Controlled Adaptive Modulation/TDMA/TDD System for High-Bit Rate Wireless Data Transmission," *Trans. on Vehicular Technology*, Vol.47, No.4, pp. 1134-1147, Nov. 1998.

[7] Proakis, J., "Digital Communications," *McGraw-Hill Series in Electrical and Computer Engineering*, 3rd Edition, 1995.

[8] Cho, K., Samueli, H., "A 8.75-MBaud Single-Chip Digital QAM Modulator with Frequency-Agility and Beamforming Diversity*," Proc. of the IEEE 2000 Custom Integrated Circuits Conference*, Orlando, FL, pp. 27-30, May 2000.

[9] Bertsekas, D., Gallager, R., "Data Networks," *Prentice Hall*, 2nd Edition, 1999.