On the efficiency of dynamic voltage scaling in a digital front end for a burst-based radio receiver

Lars Lundheim[1,3], Tormod Njølstad[2], Odd Tjore[2]

[1]SINTEF Telecom and Informatics, Trondheim, Norway
[2]Norwegian University of Science and Technology (NTNU), Department of Physical Electronics, Trondheim, Norway
[3]Norwegian University of Science and Technology (NTNU), Department of Telecommunication, Trondheim, Norway

Abstract

Dynamic voltage scaling is a well-known method to reduce the power consumption in a digital system with time-dependent processing demand. In the paper the use of this scheme in a burst based radio receiver is discussed. As an example, the burst format of HIPERLAN/2 is assumed, and the potential of power reduction in the front-end of such a receiver is studied. Simulation results indicate that with a buffer capacity in the order of 10 to 20 % of the MAC frame length, power saving in the order of 50-60% can be expected.

I. Introduction

More and more communication, both human-to-human, human-to-machine and machine-to-machine, is provided through wireless links and networks. More and more of this communication takes place via mobile, battery powered terminals. This makes it increasingly important to take power efficiency into consideration when receivers and transmitters are designed.

A well-established way of reducing power consumption in digital hardware is by lowering the supply voltage. This will increase the delay between registers so that the clock frequency must be lowered. To maintain processing throughput, the lowered clock frequency can be compensated by parallelizing the algorithms, or by introducing pipelining in the hardware architecture. This combined lowering of voltage and clock frequency is called voltage *scaling* [1].

In systems with time-varying processing requirements, *dynamic* voltage scaling has been suggested [2]. In periods with a large demand on throughput, a high supply voltage and clock frequency (V_{cc}, f_c) is used. When the demand is less, the voltage and clock frequency are lowered. By this strategy the average power consumption over time can be dramatically reduced as shown by Gutnink and Chandrakasan [2]. The use of a time-varying clock frequency for data processing requires buffers before and after the processor, allowing for asynchronous communication with the surrounding system components. This buffering introduces an extra latency or data delay which may or may not be of significance to the application at hand.

In the present paper we study the use of dynamic voltage scaling in the digital front end of a certain class of radio receivers. The basic assumption is that data are transmitted burst-wise. By burst-wise we mean that some kind of time-division duplex (TDD) and/or time division multiple access (TDMA) scheme is used, such that the receiver only needs to process certain bursts of the continuously present received radio signal. In such systems the length of each burst and the interval between the bursts are often predefined (as is the case in GSM), or they may be time-varying, as in packet-switched services, such as HIPERLAN/2. For the present study the latter variety is assumed.

There are three properties of the traffic pattern in a burst based radio system that may make dynamic voltage scaling useful. First, the bursts may require different amount of processing depending of the burst contents. (Different modulation methods and coding schemes may, in principle, be used for different bursts.) Secondly, the processing load will depend on the burst size. A third factor is that with large differences in arrival time, the available time for processing will also change, allowing for lower (V_{cc}, f_c) when bursts are sparse.

In a mobile radio system, several logical channels are present in the received data stream, and, usually, these channels have different requirements with respect to latency. This means that, whereas file transfers can allow large latencies, media access control data, such as information about when future bursts will start and stop, must be processed immediately. Conversational services, such as traditional speech telephony, fall in a class between the two extremes. A realistic assumption, therefore, is that if voltage scaling is to be used at all, the receiver must be able to distinguish between different classes of bursts, and use voltage scaling only for such data that can tolerate the associated latency. Below we will look at how the gain in power consumption changes with the amount of different types of bursts.

Assuming at least two different classes of bursts – one that can tolerate voltage scaling, and one that cannot – one must choose a hardware architecture capable of handling both. An important parameter of this architecture is the size of the buffer in front of the processor. Another one is the flexibility with respect to changing between the two burst types. Both aspects will be analyzed below.

II. A simplified traffic model for HIPERLAN/2

As an illustration and a case study for the more general parts to follow, we will consider a digital receiver front end for a HIPERLAN/2 mobile terminal (MT). For an introduction to HIPERLAN/2 physical layer, see [3].

For the further discussion, some assumptions on the traffic over the radio link must be assumed. Generalizing from HIPERLAN/2, we will assume that bursts are positioned relative to a so-called MAC frame structure. In HIPERLAN/2 a MAC frame lasts for 2 milliseconds. Each frame starts with a broadcast burst. This burst, sent by the access point (AP), gives all mobile terminals (MT) connected to the AP information about what will happen later in the frame. Depending on whether the AP has a multi sector antenna or not, several broadcast bursts may then follow. After the broadcast burst(s) one or more downlink (DL) bursts are transmitted by the AP. A DL burst is addressed to one particular MT, and each MT will receive at most one DL burst in each frame. When all DL bursts are sent, a time interval may be allocated for direct links between MTs. Following this phase, MTs may send uplink (UL) bursts to the AP. The MAC frame is finished by a random access phase, where MTs not already connected to the AP may request a connection.

Generalizing and simplifying the scheme a bit, we will assume a MAC frame of constant duration T_F. This frame is always introduced by a *high priority (HP) burst* of duration T_H that must be demodulated and decoded immediately, since it is assumed to contain information of the position of subsequent burst(s). Then, after a delay T_{HL}, a *low priority (LP) burst* of duration T_L arrives. This burst contains information that can tolerate the extra latency introduced by dynamic voltage scaling.

Note that all values above, except T_F, are stochastic variables that may change from one frame to another. For further analysis, we will assume that the resulting time series can be regarded as stationary with known mean values and variances.

III. Processor architectures

HIPERLAN/2 uses 52 out of 64 OFDM subcarriers, that are demodulated using a 64 point FFT when sampled at 20 MHz (complex samples).

Figure 1 shows a digital front end for a HIPERLAN/2 receiver. It is shown how buffers and switches are introduced around the preprocessor (PREP).

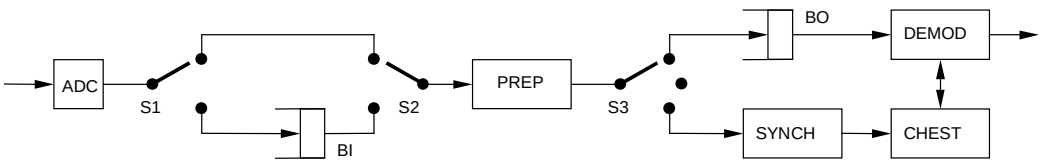


Figure 1 Architecture with provisions for dealing with two priority classes of bursts.

The figure also shows the demodulator (DEMOD), synchronizer (SYNCH) and channel estimator (CHEST). Synchronization and channel estimation is based on preambles in front of each burst.

To save power it is assumed that the ADC is switched off between bursts. Due to the MAC frame structure, the MT will always have enough information to decide when to turn it on and off. By appropriate opening and closing of the switches, the PREP can be run at full speed during reception of HP bursts, whereas a lower speed can be used for low priority data, read from the buffer BI.

Now the further analysis depends on how the hardware of the PREP is organized. We will distinguish between two main classes. A processor of the first category possesses a single set of resources. This means that two bursts can never be processed in parallel. Furthermore, it is not possible to interrupt an LP burst at the arrival of a HP burst and resume the processing when the HP burst is finished. This means that for this processor class, low-priority processing must be finished *before* the arrival of a new HP burst.

Processors of the second category are either equipped so as to be able to run two bursts in parallel, or with a double set of memory cells, so that low priority processing can be temporarily interrupted when a high priority burst arrives.

Another important aspect is the available input buffer size. Either the buffer size is so large compared to the appropriate buffer lengths that it can be considered infinite, or it is limited. Accordingly, we will considered four possible scenarios, as shown in Table 1.

Table 1 Architecture scenarios

Buffer size	**Processing across frame boundaries allowed**	**No processing across frame boundaries**
Unlimited	Scenario I	Scenario II
Limited	Scenario III	Scenario IV

IV. Performance analysis

For a given implementation of a given algorithm, the energy used by the processor for processing a burst is assumed to be proportional to the number of samples of the burst and to the square of the supply voltage V_{cc}. Assuming constant sampling frequency, the processing energy for a burst of duration T will then be expressed as $E = \gamma T V_{cc}^2$ where γ is a constant dependent on the implementation and the technology used.

For evaluating the minimum clock frequency, we will assume a known model for the maximal combinatorial delay in the processor implementation. This delay should be a function of the supply voltage expressed by $\tau(V_{cc})$. The theoretical minimum clock frequency will then be given by another function $\varphi(V_{cc}) = 1/\tau(V_{cc})$. It will be assumed that this is a strictly monotonous function, so that the inverse $\varphi^{-1}(\cdot)$ exists.

We will now study the potential for power saving by using two different supply voltages and two different clock frequencies for processing LP and HP bursts respectively. First, let us assume that a device is designed that is able to process an HP burst at the rate at which data are delivered by the ADC. Using a sampling frequency four times as large as the IF, the PREP can be fed two and two samples (I and Q) from the ADC, resulting in a an input data rate $f_S/2$. We will also assume that this is the operating clock frequency of the PREP hardware when processing an HP burst, i.e. $f_H = f_S / 2$

It is now easily shown that the ratio between the energy consumption in a PREP designed *with* voltage scaling and a design *without* voltage scaling can be expressed as

$$\rho = \frac{\tau_L \left[V_H^{-1} \varphi^{-1} \left(\frac{\tau_L f_H}{1-\tau_H} \right) \right]^2 + \tau_H}{\tau_L + \tau_H}, \qquad (1)$$

where we have defined the relative burst durations as $\tau_H = T_H / T_F$ and $\tau_L = T_L / T_F$. A value of ρ equal to 0.4, say, means that the PREP will use only 40% of the power when applying the power scaling strategy compared to no power scaling

When the function $\varphi^{-1}(\cdot)$ is known the formula (1) can now be used to calculate the power saving potential for various values of τ_H and τ_L.

V. Numerical examples

The model developed above will now be used for a given physical process for implementation, and the power saving potential will be evaluated with various assumptions.

Assuming the AMS 0.8 μm process, the combinatorial delay for one inverter with supply voltage *V* has been found to obey the expression [4]

$$\tau = K \frac{V}{(V - V_T)^\alpha}, \qquad (2)$$

where V_T = 0.9 volt is the threshold voltage, α = 1.53 is a dimensionless constant, and *K* = $352.7 \cdot 10^{-12}$ s.

For a synchronous digital circuit the clock frequency is limited by the inverse of the largest (combinatorial) delay between two registers. This delay will depend on the supply voltage and have a behaviour similar to (2), except that the constant *K* will have another value, say $1/\xi$, which can be determined experimentally for a given design. We then find that the maximum clock frequency can be expressed by

$$f_c = \varphi(V) = \xi \frac{(V - V_T)^\alpha}{V}. \qquad (3)$$

The value of ξ will now depend on the details of the digital design.

By using numerical inversion of the formula (3) one obtains power power saving potential functions as shown in Figure 2.

In principle, the relative burst durations τ_H and τ_L may have any value in the range 0 to 1. In most situations, where more than one user is involved, and communication takes place in both directions, it is assumed that both will be significantly less than 0.5**.** To show the principal behaviour, ρ is plotted through the full range of τ_L in the figure.

Curves are shown for a set of four values of τ_H . All curves are computed assuming V_H = 3.5 volt at f_H= 40 MHz. Curves are shown both for V_L restricted to the set {1.5 2 2.5 3 3.5} and assuming that V_L may take a continuum of values between 1 volt and V_H.

For moderate (and probably realistic) values of τ_H , we see that power reductions to levels in the order of 30-50% of the unscaled implementation are possible within a rather large range of values of τ_L . We see that this result holds both with an unrestrictied (continuous) voltage set, and for the restricted situation, even though the power saving is somewhat reduced in the latter case.

The results above are derived for *one* frame. The power saving potential depends on the parameters τ_H and τ_L , which change from frame to frame. We will now consider the more interesting *average* power saving potential over a large number of frames. Assuming stationary conditions with known mean values $\overline{T_L}$ and $\overline{T_H}$, we will look at the situations for the four scenarios described above.

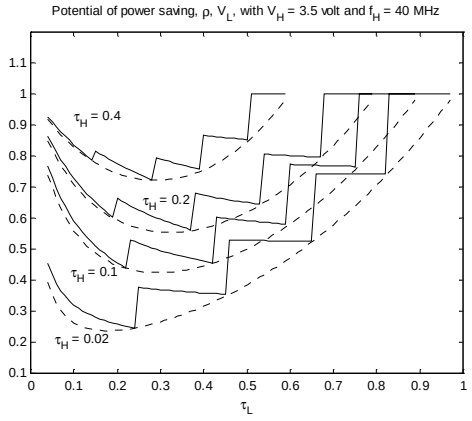


Figure 2 Power saving potential as a function of relative low priority burst duration τ_L for different values of relative high priority burst duration τ_H. (Dotted lines: Continuum of voltages available. Solid lines: Restricted set of voltages)

Scenario I

This is the most unrestricted scenario of the four. Processing load can now be distributed over several frames, and the objective of the voltage/rate control will be to minimize the total dissipated energy over the total number (or a large number) of frames. It is easily argued (using a convexity argument and Jensen's inequality in a similar fashion as done by Gutnik and Chandrakasan in [2]) that the optimal strategy in this case is to keep the clock frequency at a constant rate corresponding to the mean throughput. This means that the clock frequency f_L for low-priority bursts is given by

$$f_L = f_{L0} = f_H \frac{\overline{T_L}}{\overline{T_{LL}}}, \qquad (4)$$

where f_H is the input data rate (equal to the high priority rate), $\overline{T_L}$ is the mean duration of an LP burst, and $\overline{T_{LL}}$ is the mean available processing time between LP bursts. In the case where hardware is shared between high priority and low priority bursts, the available time is given as the difference between the frame duration and the mean duration of high priority bursts: $\overline{T_{LL}} = T_F - \overline{T_H}$. With separate hardware, $\overline{T_{LL}}$ is equal to T_F.

It is worth noting that for a stationary situation, with known mean values $\overline{T_L}$ and $\overline{T_H}$, this scenario gives a theoretical optimal performance with respect to power efficiency. This means that one can find the theoretically attainable power saving by using the power saving potential function derived above, and substituting $\overline{T_L}/T_F$ and $\overline{T_H}/T_F$ for τ_L and τ_H respectively.

Scenario II:

In this scenario the LP burst must be finished before the arrival of the next HP burst. The available time will depend on how far out in the MAC frame the LP burst arrives. To avoid this dependency, and to be able to make use of the time *between* the HP and LP bursts of a MAC frame, it is advantageous to save the LP burst in a buffer and process it during the *next* MAC frame. In this way, the average available time for processing is $\overline{T_{LL}} = T_F - \overline{T_H}$. Further analysis or simulations of this scenario has not yet been made. If the variances of T_L and T_H are small, the clock frequency can be kept close to the optimal value f_{L0} and a performance close to Scenario I can be obtained.

Scenario III:

Assuming a fixed data wordlength, the buffer size can be measured in number of samples. For our purpose it is convenient to express this quantity relative to the clock frequency at the buffer input, and express the resulting number relative to the frame length. Thus, a buffer size τ_B = 0.1, say, is capable of storing 10% of the samples belonging to a full frame.

As mentioned above, the ideal strategy for voltage/clock control is to keep these parameters constant at an operating point corresponding to the clock frequency f_{L0} defined in (4). As long as the buffer does not overflow, this is still the optimal strategy in Scenario III. (Assuming that subsequent modules can be switched off, we will not care about buffer underflow in this analysis.) To avoid overflow a time varying rate must be used in practice.

A heuristic algorithm aiming at keeping the clock frequency as close to the optimal value as possible was tried out. Different clock frequencies were used *before*, *during* and *after* reception of the LP burst. Simulations of the resulting power saving potential are shown in Figure 3 below.

From the figures we see that by using a buffer size in the order of $0.2T_F$, results close to the theoretical optimum are obtained.

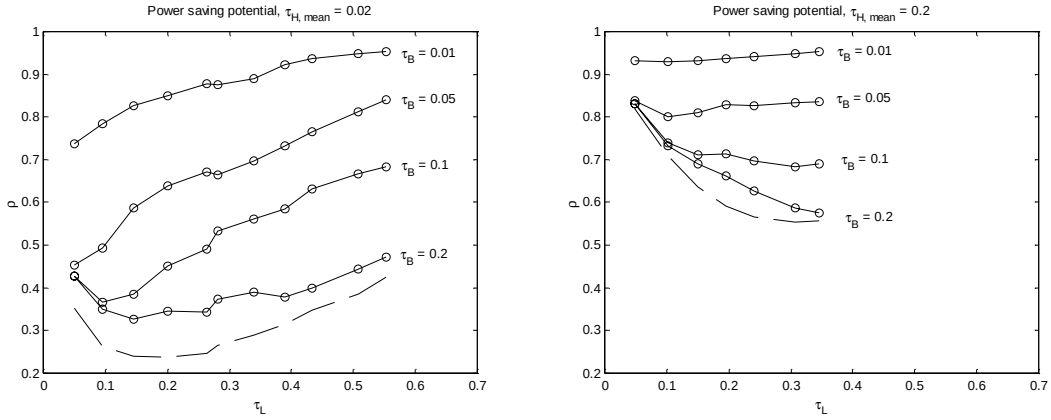


Figure 3 Mean power saving potential ρ using Scenario III for $\overline{\tau_H}$ = 0.02 (left) and $\overline{\tau_H}$ = 0.2 (right). Dashed line shows theoretical obtainable value (using Scenario I).

Scenario IV:
This is the most restricted one of the four scenarios. For small bursts, it is well approximated by Scenario I as long as the burst does not arrive too late in the MAC frame. If the buffer length is significantly smaller than the burst length, or the burs arrives late in the MAC frame the clock will have to be run at or close to f_H most of the time, and little will be gained in power efficiency. For practical cases, this Scenario is probably not very useful.

VI. Conclusions

We have expressed the potentially obtainable power saving potential in a burst-based receiver front end as a function ρ. This limit can be attained if an unlimited buffer is available in front of the processor, and if low-priority burst processing can be extended over frame boundaries (Scenario I). If, for technological reasons, the latter condition is not fulfilled (Scenario II), operation close to the theoretical is still possible. On the other hand, if buffer limitations are imposed (Scenario III), a buffer size in the order of 10 to 20 % of the MAC frame length is required

The results for Scenario II and III depend on the statistics of τ_L and τ_H, and could be refined if a more accurate traffic model were available.

VII. Acknowledgement

This work has been supported by the Norwegian research council under the program *Grunnleggende teleforskning*.

References

[1] A. Chandrakasan and R.W. Brodersen. Minimizing power consumption in digital CMOS circuits. *Proceedings of the IEEE*, 83(4):498-523, 1995

[2] V. Gutnik and A. P. Chandrakasan: "Embedded Power Supply for Low-Power DSP*", IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 5, NO. 4, December 1997, pp. 425-435.

[3] ETSI TS 101 475 (V.1.1.1): "Broadband Radio Access networks (BRAN); HIPERLAN Type 2; Physical (PHY) layer".

[4] T. Njølstad and E.J. Aas. Validation of an accurate and simple delay model, and its application to voltage scaling. Proc. of IEEE 1998 International Symposium on Circuits and Systems (ISCAS'98*),*