Area-Power Trade-Offs for Flexible Filtering in Green Radios

Navin Michael, Christophe Moy, Achutavarrier Prasad Vinod, and Jacques Palicot

Abstract: **The energy efficiency of wireless infrastructure and terminals has been drawing renewed attention of late, due to their significant environmental cost. Emerging green communication paradigms such as cognitive radios, are also imposing the additional requirement of flexibility. This dual requirement of energy efficiency and flexibility poses new design challenges for implementing radio functional blocks. This paper focuses on the area vs. power trade-offs for the type of channel filters that are required in the digital frontend of a flexible, energy-efficient radio. In traditional CMOS circuits, increased area was traded for reduced dynamic power consumption. With leakage power emerging as the dominant mode of power consumption in nanoscale CMOS, these trade-offs must be revisited due to the strong correlation between area and leakage power. The current work discusses how the increased timing slacks obtained by increasing the parallelism can be exploited for overall power reduction even in nanoscale circuits. In this context the paper introduces the notion of 'area efficiency' and a metric for evaluating it. The proposed metric has also been used to compare the area efficiencies of different classes of time-shared filters.**

Index Terms: **Channelization, cognitive radios, finite impulse response (FIR) filters, green radios, nanoscale complementary metal oxide semiconductor (CMOS).**

I. INTRODUCTION

There is a growing awareness and consensus about the fact that, the issues of energy efficiency and climate change are closely intertwined. Hence, tackling the latter problem cannot be achieved without first resolving the issue of energy efficiency. The information and communication technologies infrastructure currently contributes to nearly 3% of the total energy consumption, and the contribution is expected to double every five years [1]. Mobile telephony in particular is emerging as one of the major areas of focus for reducing the overall energy budget, due to the exponential rate at which the wireless industry has been increasing its customer base. These observations are stimulating an interest in 'green radios,' which is the name given to the growing body of research on energy efficiency in wireless networks [2]. The operational carbon footprint of the cellular infrastructure is much higher than that of mobile terminals [2], [3]. For mobile terminals, the energy efficiency is strongly linked to usability and battery life. From an environmental perspective, energy efficiency in terminals is important, because of the serious problem posed by toxic elements in the discarded batteries [4]. The sheer number of mobile terminals can translate to a significant amount of battery pollution. The primary focus in developing the wireless infrastructure has always been on spectral efficiency and capacity, not on energy efficiency. Minimizing the energy consumption of the entire wireless network requires a concerted effort at the various levels of the networking hierarchy, from low-power hardware in the infrastructure and terminals, to novel energy-aware communication paradigms.

The demand for increased bandwidth and value added wireless services is insatiable. A frequency spectrum is a scarce resource, and the need for increasing spectral efficiency necessitates the use of signal processing algorithms of dramatically increased algorithmic complexity [5]. The increase in algorithmic complexity has been accompanied by a growing demand for flexibility. New communication paradigms such as cognitive radios [6] and symbiotic networks [7] use a cooperative spectrum-sharing model to make more efficient use of the available spectrum. These paradigms also require the underlying hardware to operate in a heterogeneous wireless environment, which necessitates that the hardware be highly flexible. Flexibility along with spectrum awareness can be used for jointly targeting both spectrum and energy efficiency [8], [9]; hence, these paradigms can be considered within the ambit of green radios.

From the perspective of radio hardware design, there has been an increasing trend to push a significant portion of the radio signal processing load into the digital domain. The digital domain has the advantages of easier reprogrammability and more mature power optimization strategies. These trends suggest that the power consumption of functional blocks in the digital baseband is an important target for system-level power optimization of wireless infrastructure and terminals. This paper focuses on the problem of flexible filtering, which is required for channel selection in the digital frontend of flexible radio terminals. Flexibility necessitates the use of generic programmable multiply and accumulate (MAC) units, which have a large area overhead. Hence, one of the important architectural-level design parameters is the number of MAC units that must be instantiated in the design. At one extreme, each filter tap can be associated with its own dedicated hardware multiplier; at the other extreme, all the MAC operations can be multiplexed on to a single MAC unit. The optimum parallelism vs. power tradeoff is strongly coupled to the underlying technology. The complementary metal oxide semiconductor (CMOS) has been the technology of choice for implementing integrated circuits in the digital baseband. In higher device geometries, the power consumption of CMOS technologies

Manuscript received October 1, 2009.

This work was supported in part by the Merlion PhD Grant, under the France-Singapore Cooperation Platform for Science and Technology.

N. Michael and A. P. Vinod are with the School of Computer Engineering, Nanyang Technological University, Singapore, email: {navi0001, asvinod}@ntu.edu.sg.

C. Moy and J. Palicot are with the École Supérieure d'Électricité (SUPELEC) /Institut d'Électronique et de Télécommunications de Rennes (IETR), Cesson Sevigne, France, email: {chrisophe.moy, Jacques.palicot}@supelec.fr.

1229-2370/10/$10.00 © 2010 KICS

was dominated by dynamic or switching power, with a quadratic dependence on the supply voltage. Parallelism is generally used to trade increased area for increased timing slacks in the critical path. In higher device geometries, the relaxed performance requirements with increased parallelism was exploited for lowering the supply and threshold voltages. This resulted in a slower circuit with a significantly lower level dynamic power. However, leakage power has started overtaking the dynamic power component in nanoscale CMOS technologies. The leakage power consumption is strongly correlated to the total number of leaking transistors [10], and hence, the area. This correlation requires the traditional architectural level area-power trade-offs and the viability of parallelism as a power reduction strategy to be reexamined for smaller geometries. The contribution of this paper is twofold. Firstly, it offers some insights, into how the timing slack offered by increased parallelism can be used to reduce the overall power consumption, even in nanoscale technologies. Secondly it introduces the notion of 'area efficiency' which is a measure of the efficiency with which area can be traded for timing slack in a fixed throughput system. The notion of area efficiency is illustrated with a case study of alternate implementations of finite impulse response (FIR) filters, which are required for flexible channel selection in the digital frontend.

The paper is organized as follows: Section II highlights the need for flexibility in the channel filters of emerging communication paradigms. Section III gives an overview of the CMOS power consumption components and their dependency on various device-level and circuit-level parameters. Section IV shows how a trade-off between area and increased timing slack can be used reducing the individual power consumption components. Section V introduces the notion of area efficiency and demonstrates it with different classes of FIR filters. Section VI offers the conclusion and proposed future work.

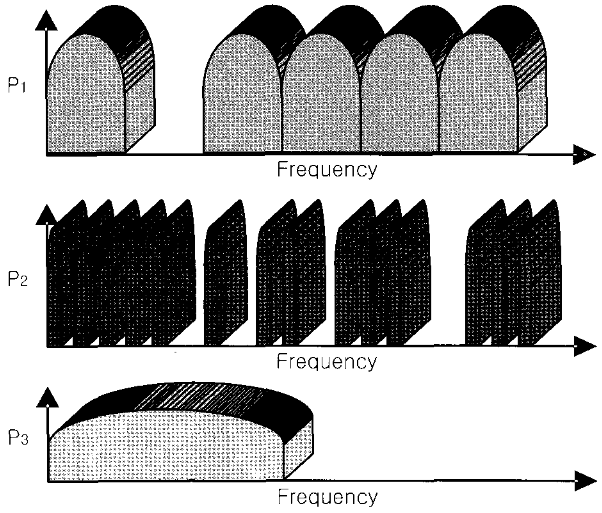

Fig. 1. Spectrum sensing scenario in cognitive radio.

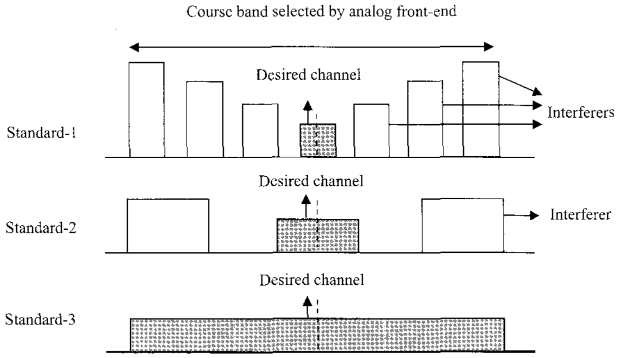

Fig. 2. Multistandard channelization.

II. FLEXIBLE FILTERING FOR CHANNEL SELECTION IN EMERGING COMMUNICATION PARADIGMS

Spectrum aware communication paradigms like cognitive radios originally aimed to increase spectral capacity by making more efficient use of the spectrum; however knowledge of the spectrum can also be translated into reduced energy consumption [8], [9]. The cognitive radio equipment needs to sense the spectrum over multiple bands, in order to detect the presence of standards with varying channel bandwidth. It uses vacant channels in a shared spectrum to receive and transmit data. Hence the bandwidth and interference attenuation requirement of the channel selection filters can vary over time. This type of scenario is illustrated in Fig. 1, where P_1, P_2, and P_3 indicate three different sensing periods. The radio analyzes the spectrum for channels of different bandwidths in the different sensing periods. Besides opportunistic spectrum sharing, similar requirements of flexibility exist in paradigms such as 4G [11], which require the radio to operate in a heterogeneous, multistandard environment. This requirement in turn imposes a requirement of flexibility on the channel selection filters and filter banks. One potential usage scenario of a flexible, spectrum-aware/energy-aware radio, could be a vertical handover from an universal mobile telecommunications system (UMTS) network to a global system for mobile communication (GSM) network, whenever free GSM channels are available. This could reduce the energy consumption of the battery.

Channel filters extract the channel of interest from a digitized wideband signal from an analog-to-digital converter (ADC). As shown in Fig. 2, pushing the fine channel selection load to the digital domain results in a very coarse band being selected by analog frontend.

The output of the ADC may comprise of unwanted interferers, blockers, and quantization noise, all of which must be removed by the channel filter, before the signal can be fed to the subsequent baseband blocks. In a multistandard system, the channel bandwidth, the interferer location and the interferer power levels can vary over time. Hence, the coefficient set and the filter lengths must be variable to provide multistandard support. Channel selection is traditionally performed by FIR filters due to the relative ease of providing a linear phase and stability. FIR channel filters are computationally intensive, due to the large number of MAC operations required. The sheer number of MAC operations per second places it beyond the computational capabilities of generic digital signal processors. Instead, the function must be performed with a dedicated hardware accelerator. Accelerators use a spatial computation style, whereas processors use a temporal computation style. The requirements of flexibility and reprogrammability, however, preclude the use of constant coefficient filter optimizations such as common sub expression elimination and graph dependency algorithms, that are traditionally used to reduce the complexity of spatial implementations of a constant coefficient FIR filter. These behavioral optimizations significantly reduce the arithmetic complexity by reducing the multiply operations to a set of shift and add operations [12], [13]. Programmability necessitates the use of generic programmable MAC units.

Because generic MAC units have a significant area overhead, the FIR filter can be implemented as a time-shared filter, where multiple coefficient multiplications are mapped to each MAC unit [14], [15]. Time-shared filters are more flexible because they allow multiple filter lengths to be folded onto the same folding set. The use of a random access memory or a register file for storing the coefficients in the time-shared filter, allows them to be updated at runtime. The control logic is usually a simple counter, whose control word can be updated for supporting multiple filter lengths.

In any particular mode of operation, the throughput requirements of the channel filter are fixed. In such a scenario, the area overhead due to the increased number of MAC units can be traded for the reduced frequency of operation and increased timing slack in the critical paths. In higher geometries supply and threshold voltage were lowered for exploiting the timing slacks and reducing the dynamic power consumption. However, lowering the threshold voltage has the effect of exponentially increasing the subthreshold leakage power, which is one of the dominant power consumption mechanisms in nanoscale CMOS circuits. Parallelism results in increased area and increased number of leaking transistors. Hence it remains to be seen whether, the relaxed performance requirements of the MAC units with increased parallelism can be translated into lower power consumption in nanoscale CMOS technologies.

III. POWER CONSUMPTION IN NANOSCALE CMOS TECHNOLOGIES

The evolution of CMOS technologies has more or less followed Moore's law which predicts that the density of transistors would double, every 24 months [16]. For that to happen, the device dimensions would need to be scaled by a factor of 0.7 in each successive generation. In the constant field scaling regime, the lateral and vertical device dimensions, the supply voltage and the threshold voltage are all scaled by a factor of 0.7, whereas the doping density is scaled by a factor of 1/0.7. This degree of scaling enables the magnitude of the electric field in the gate oxide and the MOSFET inversion channel to remain more or less unchanged. The main attraction of technology scaling lies in the fact that both the delay and the capacitance are scaled by a factor of 0.7. Hence the maximum operating frequency of the transistors is increased by about 43% while the scaled capacitance and supply voltage reduce the switching energy by a factor of 0.73 for each generation [17].

For more than three decades, aggressive constant field scaling has led to increased performance, reduced dynamic power consumption and increased transistor density. However, continued scaling has become unsustainable in the nanoscale era due to the increased role of leakage power. In higher device geometries, leakage was considered negligible in comparison to dynamic power, and was an important consideration only for portable battery-powered devices, which spent a significant amount of time in the standby mode. Active power consumption was always dominated by dynamic power. However, the leakage power contribution exceeds the dynamic power consumption in the sub-65nm geometries. This behavior can be attributed to two main factors. Firstly, scaling of the supply voltages means that the threshold voltage must also be scaled down to maintain the gate overdrive. Reducing the threshold voltage causes an exponential increase in the subthreshold leakage current [18]. Secondly, scaling of the silicon dioxide gate oxide causes the gate and the channel to be separated by the thickness of just a few atoms, and this effect tremendously increases the gate tunneling currents. The cumulative effect of these two currents has resulted in the leakage power overtaking the dynamic power as the dominant source of power consumption in nanoscale CMOS circuits. Fig. 3 shows the relative contribution of the major power consumption components over the years [10]. The nanoscale CMOS power consumption components are discussed in more detail below.

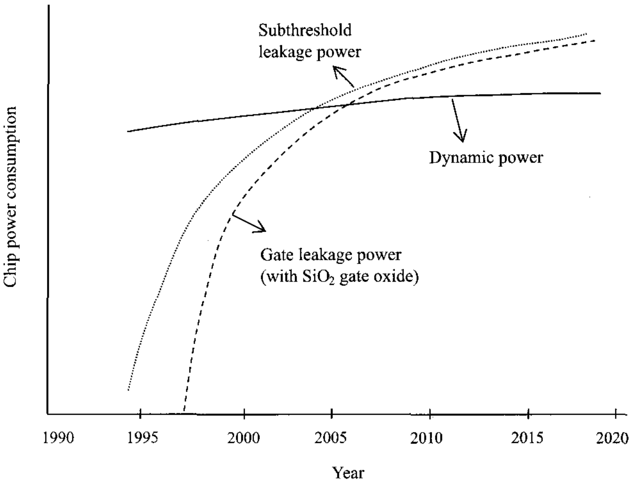

Fig. 3. CMOS power consumption trends.

A. Subthreshold Leakage Power

The subthreshold leakage current, shown in Fig. 4, refers to the drain-to-source current that flows when the gate-to-source voltage, V_{GS}, of the transistor is below the threshold voltage. Under this condition, the transistor is said to be in a weak inversion mode. The current in the weak inversion mode is dominated by the diffusion current, rather than the drift current which dominates the drain-to-source current in a strong inversion mode. This behavior can be attributed to the low concentration of minority carriers and the smaller longitudinal electric fields.

In long channel transistors, electrons are attracted to the channel from a highly doped source, under the effect of a positive gate voltage. With increased scaling of the channel length, the drain and source depletion regions are very close to each

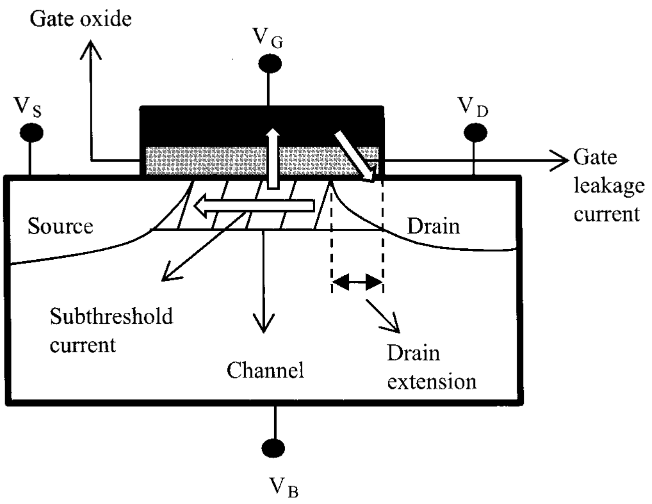

Fig. 4. Leakage currents in a MOSFET.

other. This close proximity causes the electrons in the source to be pulled into the channel under the effect of the drain electric field. Hence, in short channel transistors, the drain bias, V_{DS}, influences the channel formation in as well as the gate voltage. The resultant lowering of the potential barrier for electrons in the source to enter the channel, reduces the threshold voltage, and the effect is called drain-induced barrier lowering (DIBL) [18].

Another important factor that affects the threshold voltage is the source-body bias. For a MOSFET to function properly, both the source-body p-n junctions and the drain-body p-n junctions have to be reverse biased. Electrons are attracted to the channel from the source, whenever the channel depletion width is comparable to that of the source-body depletion width. At this point, the electrostatic potential of the channel surface is equal to that of the source region. Increasing the magnitude of the reverse-biased source-to-body voltage V_{SB}, results in the widening of the source-body depletion region. Consequently, an increased gate voltage must be applied to further increase the channel depletion width. The dependence of the threshold voltage, V_{th}, on the body bias, V_{SB}, is modeled as follows [18]:

$$V_{th} = V_{FB} + 2\psi_B + \frac{\sqrt{2\epsilon_{si} q N_A (2\psi_B + V_{FB})}}{C_{ox}} \qquad (1)$$

where V_{FB} is the flatband voltage, N_A is the doping density, C_{ox} is the gate oxide capacitance per unit area, and ψ_B is the difference between the Fermi potential and substrate intrinsic potential. The cumulative effect of DIBL and body bias can be incorporated into (1) as follows [18]:

$$V_{th} = V_{FB} + 2\psi_B + \frac{\sqrt{2\epsilon_{si} q N_A (2\psi_B + V_{FB})}}{C_{ox}} - \eta V_{DS} \qquad (2)$$

where η is the DIBL coefficient. When the body bias and DIBL are taken into consideration, the substrate leakage power can be modeled as follows [18], [19]:

$$P_{sub} = D V_{DD} \exp\left(\frac{V_{GS} - V_{th}}{m v_T}\right)\left(1 - \exp\left(\frac{-V_{DS}}{v_T}\right)\right) \qquad (3)$$

where D is given as:

$$D = \mu_0 C_{ox} \frac{W}{L}(m-1) v_T^2. \qquad (4)$$

V_{DD} is the supply voltage, m is the body effect coefficient, v_T is the thermal voltage, μ_0 is the zero bias mobility, W is the gate width and L is the gate length. The body effect coefficient, m, is expressed as (5):

$$m = 1 + \frac{3T_{ox}}{W_{dm}} \qquad (5)$$

where W_{dm} is the maximum depletion layer width, while T_{ox} is the SiO$_2$ gate oxide thickness. The equation (3) indicates an exponential relation between the subthreshold leakage power and the threshold voltage.

B. Gate Leakage Power

In an n-channel transistor, the n+ doped polycrystalline silicon gate is separated from the n-type channel by an insulating layer of silicon dioxide (SiO$_2$), which is assumed to have infinite impedance. However, aggressive scaling has resulted in a gate oxide approaching a thickness of approximately 1.2 nm or, equivalently, 5 silicon atoms. At this thickness, the electrons can directly tunnel through an extremely narrow potential barrier, resulting in gate leakage currents [18], [20]. A reverse situation exists in the off state ($V_{GS} = 0$, $V_{DS} = V_{DD}$), when the electrons in the n+ polycrystalline gate tunnel directly into the drain extension regions. Fig. 4 shows both the mechanisms. Similar hole tunneling currents also exist in p-channel transistors. The power consumption due to the direct tunneling gate leakage current can be modeled as follows [21], [22]:

$$P_{gate} = A V_{DD} W L \left(\frac{V_{ox}}{T_{ox}}\right)^2 \exp\left(-B\left[\frac{1-\left(1-\frac{V_{ox}}{\phi_B}\right)^{1.5}}{\frac{V_{ox}}{T_{ox}}}\right]\right) \qquad (6)$$

where ϕ_B is the barrier height, V_{ox} is the voltage drop across the gate oxide and the constants A and B are functions of the effective carrier mass and the barrier height ϕ_B. A closed-form expression for V_{ox} in terms of the gate voltage, the Fermi level and the flatband voltage can be found in [21]. The expression (6) shows that the gate leakage power has an exponential dependence on the supply voltage and the thickness of the gate oxide.

C. Dynamic Power

Dynamic power consumption occurs when switching activity induces charging and discharging of capacitive nodes. The switching activity comprises the useful activity that is necessary for evaluating a particular output and the spurious switching activity that a node may experience, before settling into a steady state value. The load capacitance typically includes both parasitic transistor capacitance and interconnect capacitance. In scaled technologies, the latter component has started dominating due to the non-scaling of the interconnect capacitance. The dynamic power consumption can be expressed as follows [23]:

$$P_{dyn} = \alpha_f C_T V_{DD}^2 f_{clk} \qquad (7)$$

where C_T stands for the total capacitive load, α_f is the fraction of capacitive nodes that are switching, and f_{clk} is the clock frequency. The expression (7) reveals a quadratic dependence on the supply voltage.

IV. TRADING AREA FOR LOWER POWER CONSUMPTION

A. Traditional Architecture-Driven V_{DD}/V_{th} Scaling

The delay of a CMOS gate is determined by the time taken to charge and discharge the capacitive nodes, and that time depends on the saturation current. The saturation current is in turn strongly dependent on the supply voltage, V_{DD}, and the threshold voltage, V_{th}. The dependence of the gate delay on these parameters can be given by the alpha-power law MOSFET model [24] as shown in (8)

$$T_{pd} = \frac{K_D V_{DD}}{(V_{DD} - V_{th})^{\alpha}} \tag{8}$$

where K_D is a fitting parameter and α is the velocity saturation term, which is equal to 1.5 in short channel transistors [24].

In higher device geometries, dynamic power consumption is the predominant mode of active power dissipation; the other components are negligible. The quadratic dependence of supply voltage on dynamic power makes supply voltage scaling an attractive option. The reduction in the gate overdrive can be partly offset by lowering the threshold voltage which enables the supply voltage to be scaled over a wider range. This dual scaling of both supply and threshold voltages yields greater reductions in the dynamic power, than possible by scaling the supply voltage alone. Traditional architecture-driven voltage scaling techniques in a fixed throughput system use parallelism to trade area for a lower operating frequency of the datapath operators [21]. The increased timing slack produced in the critical paths by the reduced internal cycle period, increases the room available for V_{DD}/V_{th} scaling. The physical capacitance term, C_T, in (6), undergoes a linear increase due to increased parallelism. However this increase is easily compensated by the quadratic reduction that stems from the lower supply voltage.

B. Exploitation of Timing Slack in Nanoscale CMOS

This section discusses how the timing slacks produced by increased parallelism can be used to reduce the total power consumption even in nanoscale CMOS circuits. This reduction is achieved by suitably manipulating the gate oxide thickness, the body bias and the supply voltage. The traditional architecture-driven V_{DD}/V_{th} scaling model faces some serious bottlenecks in nanoscale technologies. A reduction in the supply voltage causes a quadratic reduction in the dynamic power and an exponential reduction in gate leakage; however lowering the threshold voltage causes an exponential increase in the subthreshold leakage current. Both the gate leakage and the subthreshold leakage are also linearly dependent on the gate width. Hence the total leakage power consumption is strongly correlated with the total gate width and area [10]. The total physical capacitance term, C_T, in (6) is also strongly correlated with the total area. Hence, the power reduction strategy should ensure that the potential gains from the increased parallelism are not offset by the area overheads.

Expression (5) indicates that an exponential reduction in gate leakage power can be obtained by either reducing V_{DD} or increasing T_{ox}. The gate oxide capacitance per unit area, C_{ox}, can be expressed in terms of T_{ox} as follows:

$$C_{ox} = \varepsilon_{ox}/T_{ox} \tag{9}$$

where ε_{ox} is the permittivity of the gate oxide. The expression for the threshold voltage in (2), can be reformulated using (9) as shown below.

$$V_{th} = V' + \frac{T_{ox}\sqrt{2\epsilon_{si} q N_A (2\psi_B + V_{FB})}}{\varepsilon_{ox}} \tag{10}$$

where

$$V' = V_{FB} + 2\psi_B - \eta V_{DS}. \tag{11}$$

When T_{ox} is increased, the threshold voltage increases, resulting in an exponential reduction in subthreshold leakage. Increasing the gate oxide thickness reduces the gate capacitance, which in turn reduces the dynamic power consumption. Increased T_{ox} also causes an exponential redcution in the gate leakage power.

The underlying idea of using thicker gate oxides to reduce the leakage currents has been used in a multiple oxide CMOS process [25]. In that process, the non-critical paths are implemented using low performance, high T_{ox} transistors whereas the critical paths are implemented by using high speed transistors with a lower T_{ox}. When the number of critical paths is much lower than that of the non-critical paths, this technique significantly reduces the leakage power consumption. However, the drawback of the technique is that, it requires a complicated technological process for fabrication. The threshold voltage, V_{th}, increases linearly with T_{ox} as indicated in (10). Increasing the oxide thickness has the unwanted side-effect of increasing the short channel effects caused by the reduced aspect ratio. The aspect ratio is directly proportional to the channel length, L, and inversely proportional to $T_{ox}^{1/3}$ [18]. Hence, maintaining the aspect ratio requires an increase in the length of the channel, when the thickness of the gate oxide is increased. When supply voltages are also simultaneously scaled down, the room available for varying the oxide thickness is however limited due to the rapidly increasing gate delay, as indicated by (8) and (10). As discussed below, this problem can be redressed by a combination of architectural level and device level power optimization strategies.

Note that the expression for the threshold voltage is also a function of the body bias. In n-channel transistors, reverse body bias occurs when $V_{SB} > 0$. Reverse-body bias has been used as a standby leakage reduction strategy for fast transistors with a high leakage current. The increased source-body depletion width under this condition, yields an increased threshold voltage and, hence, reduced subthreshold current in standby modes. Removing the reverse body bias allows the circuit to have a high performance in the active mode. However, this strategy provides diminishing returns in scaled technologies in which, a higher gate oxide capacitance C_{ox} and lower doping density N_a have resulted in a weakened body effect. Reverse body bias also results in increasing the drain-body depletion width, aggravating the short channel effects which in turn lower the threshold voltage [26]. This method also ensures reduction of only the subthreshold leakage, and not the gate leakage component, which is significant in aggressively scaled geometries.

A forward body bias for n-channel transistors occurs when $V_{SB} < 0$. It can be used as an alternate leakage reduction strategy in which the performance of nominally high V_{th} devices, with a low standby leakage power can be improved when a forward body bias is applied in the active mode. One important advantage of this technique is that it can scale well into lower geometries because of reduced short channel effects under the forward body biased condition [27]. The threshold voltage is also more sensitive to a fprward bias voltage. The application of a forward body bias allows the gate oxide thickness to be increased further, while maintaining a given threshold voltage. However, the magnitude of the applicable forward bias is limited to around 600 mV due to the need for maintaining a reverse bias at the drain-body and source-body p-n junctions of the MOSFET.

An increased parallelism lowers the circuit's frequency of operation and relaxes the performance requirements. The combined use of forward body bias and increased parallelism now allows the whole circuit to be implemented using a thicker gate oxide, than would have been possible with only either of the methods. Note that the focus here is to exploit power optimization strategies at both architectural and physical design levels, in order to allow the entire circuit to be implemented with a thicker gate oxide and not just the non-critical paths, as is the case with a multiple oxide process. Hence, these strategies reduce the leakage and dynamic power consumption components without having to resort to an advanced CMOS process. The lower supply voltage and higher threshold voltage produce an exponential reduction in subthreshold and gate leakage currents, and a quadratic reduction in dynamic power. These effects compensate the near linear increase in leakage power and dynamic power due to increased area.

The problem of deriving the optimum supply and threshold voltages for the purpose of minimizing the overall power can be framed as follows: Assume that, for a given degree of parallelism in a fixed throughput system, the internal cycle period of the datapath operators is T_C. Let T_D be the propagation delay of the operator for the nominal supply and threshold voltages. The available timing slack for V_{DD}/V_{th} optimization is given by $(T_C - T_D)$. If off state conditions are assumed for subthreshold leakage ($V_{GS} = 0, V_{DS} = V_{DD}, V_{DD}/v_T \to 0$), the subthreshold power consumption, can be given as:

$$P'_{sub} = DV_{DD}^2 e^{\frac{-V_{th}}{mv_T}}. \tag{12}$$

The total power consumption can now be given as:

$$P_{tot} = P'_{sub} + P_{dyn} + P_{gate}. \tag{13}$$

Let $V_{\max}$ be the maximum permissible forward bias voltage ($V_{SB} = -V_{\max}$). Substituting $V_{DS} = V_{DD}$ in (10), the gate oxide thickness, T_{ox}, can be expressed in terms of the threshold voltage as follows:

$$T_{ox} = \frac{(V_{th} - V_{FB} - 2\psi_B + \eta V_{DD})\,\varepsilon_{ox}}{\sqrt{2\varepsilon_{si} q N_A\,(2\psi_B - V_{max})}}. \tag{14}$$

The closed-form expression for V_{ox} in [21] can be used to express V_{ox} in terms of V_{DD} by assuming the on-state condition, $V_{GS} = V_{DD}$. When (5) and (13) are used, the total power consumption, P_{tot}, can be completely expressed in terms of the two variable parameters: Namely, the supply voltage, V_{DD}, and the threshold voltage, V_{th}. The minimization of the total power consumption can now be treated as a constrained optimization problem [28] in which, the objective is to find the optimum (V_{DD}, V_{th}) pair which minimizes the total power, subject to the following constraint:

$$\frac{K_D V_{DD}}{(V_{DD} - V_{th})} \le T_C - \Delta T_C \tag{15}$$

where ΔT_C is a safety margin introduced to accommodate threshold voltage variations. Additional constraints can be imposed on the basis of practical and technological limits of the supply and threshold voltage. With (14) the optimum (V_{DD}, V_{th}) pair can be used to evaluate the gate oxide thickness. The use of the maximum permissible forward bias voltage, V_{max}, ensures that the maximum possible oxide thickness is obtained for a given (V_{DD}, V_{th}) pair. The transmission protocols used by the infrastructure might use some built in standby periods, during which some of the accelerator cores may be powered down. For these intervals, the subthreshold leakage power consumption can be reduced further by removing the forward body bias ($V_{SB} = 0$).

V. AREA EFFICIENCY OF TIME-SHARED FIR FILTERS

The last section proposed a physical design strategy to minimize the total power consumption, for a given degree of parallelism and a given timing slack. The use of parallelism for increasing the available timing slack incurs an area penalty. Finding a suitable architectural-level structure, that offers a large increase in timing slack for a small increase in area is still an open problem. All the three major power components in nanoscale CMOS, show a near-linear dependence on the area. The physical design level power minimization strategy, discussed in the last section, depends strongly on the amount of timing slack that can be obtained by increasing the area. Aggressive scaling brings the supply and threshold voltages closer to each other. In Fig. 5, the normalized delay, $V_{DD}/(V_{DD} - V_{th})^{1.5}$, is plotted against the V_{th}, where $V_{DD} = 1\ V$ and $V_{DD} = 0.9\ V$. It can be seen that when the V_{th} approaches V_{DD}, the delay curve climbs steeply. This behavior implies that even a small change in V_{th}, causes a large increase in delay. The use of parallelism in this region incurs a large area penalty. Hence, one important architectural-level design parameter, for using parallelism in scaled technologies is the efficiency with which a parallel architecture trades area for timing slack.

One potential figure of merit that could be used for measuring the area efficiency of a fixed throughput system is the area-frequency (AF) product. The term 'area' here refers to the total area of the datapath and the memory; the term 'frequency' refers to their internal operating frequency of the datapath operators. In contrast to an architecture that has a higher AF product and a comparable area, an architecture with a lower area-frequency

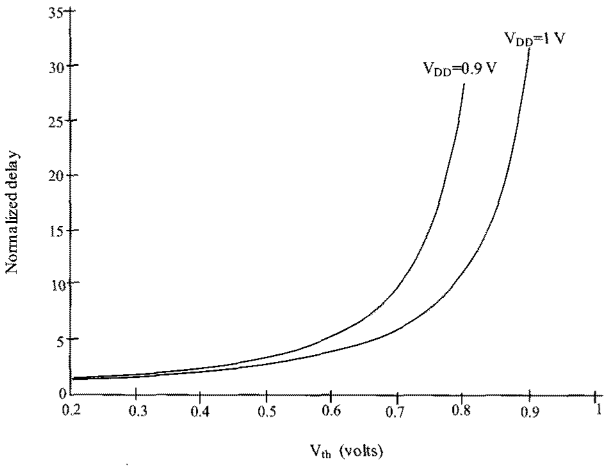

Fig. 5. Normalized delay as a function of the threshold voltage.

product has a lower operating frequency and, hence, increased cycle period. The increased cycle period helps increase the timing slack, which in turn provides more room for lowering the supply voltage and increasing the threshold voltage. The area efficiency in terms of the AF product is compared below, for different classes of time-shared filters.

A time-shared FIR filter structure can be derived by folding a direct form FIR filer structure or a transpose direct form filter structure onto a limited set of MAC units. An alternative strategy presented by us in [29], would be to use a fast filter algorithm (FFA) structure, as a starting point, for the construction the time-shared filter. FFAs [30], [31] work on the principle of algorithmic strength reduction; that is, they reduce the number of expensive MAC operations at the cost of increased add operations. The following analysis shows how the reduced number of MAC operations in FFA-based time-shared filters, can be used to reduce the operating frequency and increase the available timing slack. These results are shown to be favorable, in comparison to a folded direct form filter with a comparable area.

Let the term, T_D, be the critical path delay of a MAC unit for the nominal supply and threshold voltages. Consider a direct form filter of length N, folded onto M MAC units. M is assumed to be much smaller than N and N is assumed to be a multiple of M for simplicity. If the throughput requirement of the filter is f_{clk}, the operating frequency of each MAC can be given by Nf_{clk}/M, and the timing slack in the system can be given by $(T_D - M/Nf_{clk})$. Furthermore if the area cost of a coefficient register is A_r, the coefficient storage area overhead for each filter tap can be given by NA_r/M. Let the area cost of each MAC unit be A_m. The AF product of the folded direct form filter can be given as:

$$\begin{aligned} AF_{direct} &= M \times \left(A_m + \frac{NA_r}{M}\right) \frac{NA_r f_{clk}}{M} \\ &= A_m N f_{clk} + \frac{N^2 A_r f_{clk}}{M}. \end{aligned} \tag{16}$$

FFA structures are a special class of parallel FIR filter structures. A K-parallel FIR structure, which corresponds to an N tap FIR filter, is a block processing structure that processes K inputs in parallel and produces K outputs. This structure has K^2 subfilters, each of length N/K (where N is assumed to be a multiple of K for simplicity). When multiple inputs are processed in parallel significant redundancies exist across the subfilters. This phenomenon is used by the FFA structures to reduce the number of subfilters. A K-by-K FFA ($K \times K$ FFA) has S_K subfilters, where S_K is typically much smaller than K^2, and each subfilter is of length N/K. This reduction comes at the expense of A_K preprocessing/postprocessing adders. Fig. 6 illustrtaes the general FFA structure through an example of the simplest FFA, namely the 2×2 FFA.

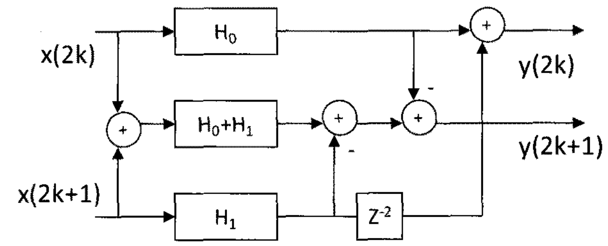

Fig. 6. 2×2 FFA.

Table 1. FFA parameters.

K	S_K	A_K
2	3	4
3	6	10
4	9	20
5	12	40
6	18	42
8	27	76

The general FFA structure consists of regular FIR subfilters and a highly irregular preprocessing/postprocessing addition network. The subfilters are either a polyphase component of the original filter, or an additive combination of different polyphase components. The time-shared filters in [29] are obtained by folding each of the regular subfilters of a base FFA filter, onto a set of L MAC units; at the same time the irregular preprocessing/postprocessing addition network data flow graph is directly mapped onto the hardware. This method exploits the regularity of the FIR subfilters for a lower control overheads; it also takes advantage of the algorithmic strength reduction of FFAs. With a throughput requirement of f_{clk}, the input sample rate for each subfilter is f_{clk}/K. Each of the subfilters is of length, N/K. If each subfilter is folded onto L MAC units, the operating rate of each MAC unit is Nf_{clk}/K^2L. The coefficient storage area overhead per filter tap in this structure can be given NA_r/KL. The preprocessing/postprocessing addition network operates at a rate of f_{clk}/K. The lower bound on the timing slack in the system and, hence, the bottleneck for V_{DD}/V_{th} scaling, is still determined by the time multiplexed MAC units. The total number of MAC units is given by $M = S_K L$. Given that the area cost of a word length adder is A_d, the AF product can be given as

$$AF_{FFA} = \left(S_K L A_m + A_K A_d + \frac{S_K N A_r}{K}\right) \frac{N f_{clk}}{K^2 L}. \tag{17}$$

Table 2. MAC, adder and register area.

Circuit	Gate count
MAC (A_m)	2098
Adder (A_d)	223
Register (A_r)	85

Table 3. Area comparison for different time-shared filters (N=36).

Filter	MAC units	Coefficient registers	Pre/Post adders	Gate count
9 MAC DFF	9	36	0	21,942
9 MAC 2×2 FFA	9	54	4	24,364
9 MAC 4×4 FFA	9	81	20	30,227

Table 1 lists the different tabulated values of S_K and A_K [31].

To estimate the values of A_m, A_d, and A_r, a 16 bit register, a 32 bit adder circuit, and a MAC circuit comprising of a 16×16 bit multiplier and a 32 bit adder-accumulator were synthesized using a TSMC 0.18 μm process. The Synopsys Design Compiler was used to estimate the cell area A_m, A_d, and A_r respectively. Table 2 shows the area in terms of the gate count. These results were obtained by normalizing the area values by the cell area of a two input NAND gate of the same library.

The area of the operators obtained above can be used to estimate the datapath and memory area of three alternative time-shared implementations of a N tap FIR filter: A direct form filter (DFF) folded onto 9 MAC units ($M = 9$), a 9 MAC-based 2×2 FFA structure ($K = 2$, $S_2 = 3$, $L = 3$), and a 9 MAC-based 4×4 FFA structure ($K = 4$, $S_4 = 9$, $L = 1$). The three examples indicate three alternative methods of increasing the parallelism of a time-shared FIR filter. Table 3 lists the estimated areas for $N = 36$. The flexibility requirement, demands that the filter length N and the throughput f_{clk} are both variable. In a comparison of relative trends, the AF product is plotted for all the three filters, against various throughput values (where N fixed at 100) in Fig. 7, and for various filter lengths (where f_{clk} fixed at 50 MHz) in Fig. 8. The plots indicate that the FFA structures have a lower AF product than the folded direct form filter, which in turn implies that the FFA structures can more effectively trade area for lower operating frequency.

The results also suggest that using a higher order FFA filter as the starting point for deriving the time-shared structure,yields a lower area-frequency product, when compared to a lower order base FFA filter with the same number of MAC units. The increased timing slack in FFA structures provides more room for minimizing the total power consumption at the physical design level because, the supply and threshold voltages can be varied over a wider range.

VI. CONCLUSION AND FUTURE WORK

This paper analyzes the role of parallelism as a power reduction strategy in nanoscale CMOS technologies and highlights the increased contribution of leakage power. Even though leakage power is strongly correlated to area, the parallelism-induced increase in timing slack can reduce both leakage and dynamic power. The proposed power reduction strategy uses the performance relaxation obtained by a combination of parallelism and the forward body bias techniques to minimize the total power consumption, without resorting to an advanced CMOS process. An architectural-level metric for fixed throughput systems, called the AF product is introduced as a means of comparing the efficiency with which alternative architectures trade area for timing slack. The results show how this metric can be used to compare different implementations of programmable time-shared FIR filters, which are an essential building block for flexible, energy efficient radios.

Designing in an AF space gives an idea about the timing slack for alternate architectures with a comparable area. However it does not give any information about the amount of power reduction achievable from the available timing slack. The degree of possible dynamic and power reduction is strongly coupled

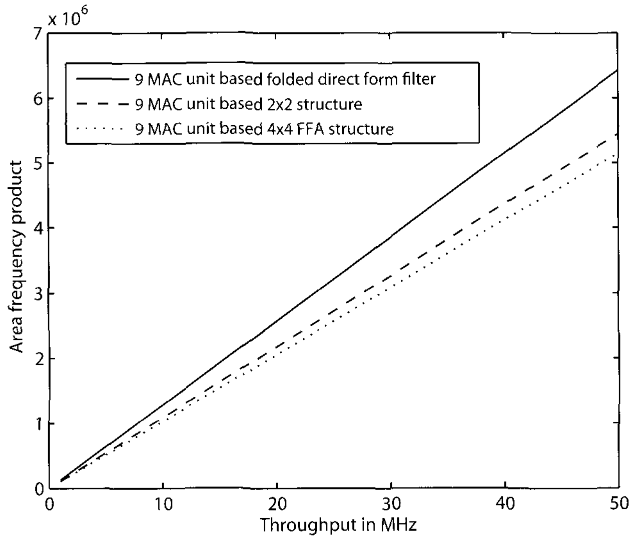

Fig. 7. AF product vs. throughput.

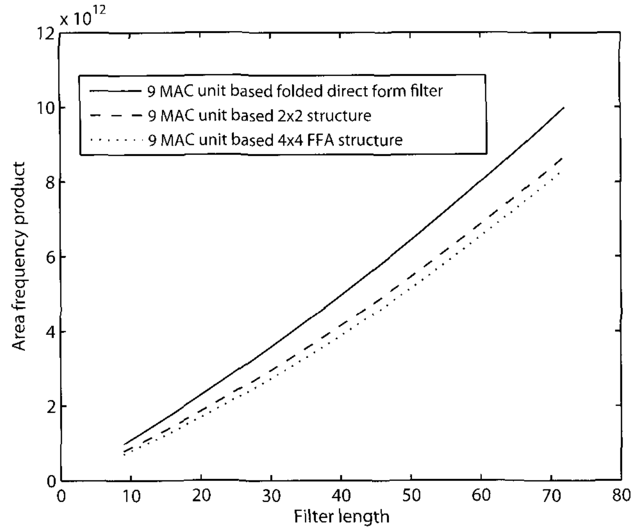

Fig. 8. AF product vs. filter length.

to nominal supply and threshold values. The leverage of parallelism is expected to decrease when scaling results in further lowering of the V_{DD}/V_{th} ratio. Hence, further research should focus on the design of fixed throughput systems in a generalized AF^n space, where the exponent n, reflects the achievable power reduction, at a specific value of supply and threshold voltage.

REFERENCES

[1] A. F. Pele, (2009, June 25). *Leti works on green mobile networks*, EE Times Europe. [Online]. Available: http://www.eetimes.eu/218101321

[2] S. Armour, T. O. Farrell, S. Flether, A. Jeffries, D. Lister, S. Mclaughlin, J. Thompson, and P. Grant, (2009, June 12). *Green radio: Sustainable wireless networks*, IET. [Online]. Available: http://kn.theiet.org/communications/green-radio-article.cfm

[3] M. Stutz, M. F. Emmenegger, R. Frischknecht, M. Guggisberg, R. Witschi, and T. Otto, "Life cycle assessment of the mobile communication system UMTS: Towards eco-efficient systems," *Int. J. Life Cycle Assessment*, vol. 11, no. 4, pp. 265–276, 2006.

[4] A. Rayapura, "Wireless waste. The challenge of cell phone and battery collection," *Inf. Report*, 2005.

[5] J. M. Rabaey, "Silicon platforms for the next generation wireless systems – What role does reconfigurable hardware play?" in *Proc. Springer-Verlag*, 2000, pp.277–285.

[6] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 201–220, Feb. 2005.

[7] R. Gedge, "Symbiotic networks," *BT Technol. J.*, vol. 21, pp. 67–73, 2003.

[8] A. He, S. Srikanteswara, J. H. Reed, C. Xuetao, W. H. Tranter, K. K. Bae, and M. Sajadieh, "Minimizing energy consumption using cognitive radio," in *Proc. IEEE Performance, Comput. and Commun. Conf.*, Dec. 2008, pp. 372–377.

[9] J. Palicot, "Cognitive radio: An enabling technology for the green radio communications concept," in *Proc. Int. Wireless Commun. and Mobile Comput. Conf.*, June 2009, pp. 21–24.

[10] N. S. Kim, T. Austin, D. Baauw, K. Flautner, J.S. Hu, M.J. Irwin, M. Kandemir, and V. Narayanan, "Leakage current: Moore's law meets static power," *Computer*, vol. 36, no. 12, pp. 68–75, Dec. 2003.

[11] J. M. Rabaey, "Wireless beyond the third generation-facing the energy challenge," in *Proc. Int. Symp. Low Power Electron. and Design*, Aug. 2001, pp. 1–3.

[12] M. Potkonjak, M. B. Srivastava, and A. P. Chandrakasan, "Multiple constant multiplications: efficient and versatile framework and algorithms for exploring common subexpression elimination," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 15, no. 2, pp. 151–165, 1996.

[13] A. G. Dempster and M. D. Macleod, "Use of minimum-adder multiplier blocks in FIR digital filters," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 42, no. 9, pp. 569–577, Sept. 1995.

[14] C. H. Wang, A. T. Erdogan, and T. Arslan, "High throughput and low power FIR filtering IP cores," in *Proc. IEEE Int. SOC Conf.*, Sept. 2004, pp. 127–130.

[15] C. Xu, C.-Y. Wang, and K. K. Parhi, "Order-configurable programmable power efficient FIR filters," in *Proc. Int. Conf. High Performance Comput.*, Dec. 1996, pp. 357–361.

[16] M. Lundstorm, "Moore's law forever," *Science*, vol. 299. no. 5604, pp. 210–211, 2003.

[17] B. Nikolic, "Design in the power-limited scaling regime," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 71–83, Jan. 2008.

[18] K. Roy, S. Mukhopadhyay, and H. M.-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[19] H. Jeon and Y. B. Kim, "A novel technique to minimize standby leakage power in nanoscale CMOS VLSI," in *Proc. IEEE Int. Instrumentation and Measurement Technol. Conf.*, May 2009, pp. 1372–1375.

[20] M. Drazdziulis, "A gate leakage reduction strategy for future CMOS circuits," in *Proc. Eur. Solid-State Circuits, Conf.*, Sept. 2003, pp. 317–320.

[21] S. P. Mohanty, V. Mukherjee, and R. Velagapudi, "Analytical modeling and reduction of direct tunneling current during behavioral synthesis of nanometer CMOS circuits," in *Proc. ACM/IEEE IWLS*, 2005, pp. 249–256.

[22] E. Kougianos and S. P. Mohanty. "Impact of gate-oxide tunneling on mixed-signal design and simulation of a nano-CMOS VCO," *J. Microelectron.*, vol. 40, no. 1, pp. 95–103, 2009.

[23] A. P. Chandrakasan and R. W. Brodersen, *Low power CMOS digital design*, Norwell, MA: Kluwer, 1996.

[24] K. A. Bowman, B. L. Austin, J. C. Eble, X. Tang, and J. D. Meindl, "A physical alpha-power law MOSFET model," in *Proc. Int. Symp. Low Power Electron. and Design*, 1999, pp. 218–222.

[25] N. Sirisantana, L. Wei, and K. Roy, "High-performance low-power CMOS circuits using multiple channel length and multiple oxide thickness," in *Proc. IEEE Int. Conf. Computer Design*, 2000, p. 227.

[26] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, "Effectiveness of reverse body bias for leakage control in scaled dual Vt CMOS IC," in *Proc. Int. Symp. Low Power Electron. and Design*, 2001.

[27] A. Keshavarzi, S. Narendra, B. Bloechel, S. Borkar, and V. De, "Forward body bias for microprocessors in 130nm technology generation and beyond," *IEEE J. Solid-State Circuits*,vol.38, pp. 696–701, 2003.

[28] M. Miyazaki, J. Kao, and A. P. Chandrakasan, "A 175mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2002, pp. 58–59. 27

[29] N. Michael, A. P. Vinod, C. Moy, and J. Palicot, "Design of low power multimode time-shared filters," in *Proc. Int. Conf. Info., Commun. and Signal Process.*, Dec. 2009, pp. 1–5.

[30] A. Parker and K. K. Parhi, "Low area/power parallel FIR digital filter implementations," *J. VLSI Signal Process.*, vol. 17, no. 1, pp. 75–92, Sept. 1997.

[31] Z. J. Mou and P. Duhamel, "Short-length FIR filters and their use in fast nonrecursive filtering," *IEEE Trans. Signal Process.*, vol. 39, pp. 1322–1332, June 1991.

Navin Michael was born on 29 May, 1985, in Chennai, India. He received the B.Tech. degree in Information and Communication Technology from DAI-ICT, Gandhinagar, India, in 2007. Since August 2007, he has been pursuing his Ph.D. from Nanyang Technological University, Singapore. His research focuses on the flexible and low power implementation of the digital front-end in multimode software defined radios. Between June 2008 to November 2008 and September 2009 to February 2010, he was working as a visiting researcher with the SCEE team at SUPELEC/IETR, Rennes, France. These visits were supported by the French Embassy of Singapore, as a part of the Merlion Ph.D. Grant 2007. His major interests are software defined radios, cognitive radios, green communications, low power signal processing hardware, and channelization.

Christophe Moy received the engineer diploma of the INSA (National Institute of Applied Sciences), Rennes, France. in 1995. He received his M.Sc. and Ph.D. degrees in Electronics in 1995 and 1999 from the INSA. He then worked 6 years at Mitsubishi Electric ITE-TCL research lab where he was focusing on Software Radio systems and concepts, including digital signal processing, HW and SW architecture, co-design methodology, and reconfiguration. He represented Mitsubishi Electric at the SDR Forum and worked on French research program A3S, and IST European project E^2R. Since 2005, he has been working as a Professor in SUPELEC. His research, which focuses on Software Radio and Cognitive Radio, is done in the IETR entity of CNRS. He addresses heterogeneous design techniques for SDR, as well as high-level design for cognitive management and decision making inside the cognitive cycle. He is participating to the IST Network of Excellence $NEWCOM++$ and SEC EULER project as well as a French ANR project on SDR design called Mopcom. He was also involved in IST projects E^2R phase 2 and NEWCOM, and French ANR project Idromel.

Achutavarrier Prasad Vinod received his B.Tech. degree in instrumentation and control engineering from University of Calicut, India in 1994 and the M.Eng. and Ph.D. degrees in computer engineering from Nanyang Technological University, Singapore in 2000 and 2004 respectively. He has spent the first 5 years of his career in industry as an automation engineer at Kirloskar, Bangalore, India, Tata Honeywell, Pune, India, and Shell Singapore. From September 2000 to September 2002, he was a lecturer in the School of Electrical and Electronic Engineering at Singapore Polytechnic, Singapore. He was a lecturer in the School of Computer Engineering at Nanyang Technological University (NTU), Singapore, from September 2002 to November 2004, and since December 2004, he has been an assistant professor in NTU. His research interests include digital signal processing (DSP), low power and reconfigurable DSP circuits, software radio, cognitive radio, and brain-computer interface. He has published 100 papers in refereed international journals and conferences. He is an editor of the International Journal of Advancements in Computing Technology and a Senior Member of IEEE.

Jacques Palicot received, in 1983, his Ph.D. degree in Signal Processing from the University of Rennes. Since 1988, he has been involved in studies about equalization techniques applied to digital transmissions and analog TV systems. Since 1991, he has been involved mainly in studies concerning the digital communications area and automatic measurements techniques. He has taken an active part in various international bodies, such as EBU, CCIR, URSI, and within the RACE, ACTS, and IST European projects. He has published various scientific articles notably on equalization techniques, echo cancellation, hierarchical modulations and Software Radio techniques. He is currently involved in adaptive signal processing and in new techniques, such as software radio and cognitive radio. From November 2001 to September 2003, he had a temporary position within INRIA/IRISA in Rennes. Since October 2003, he has been with SUPELEC in Rennes where he leads the Signal Communications and Embedded Electronics (SCEE) research team.