Citation: Beckett, P 2005, 'Low-power spatial computing using dynamic threshold devices', in IEEE International Symposium on Circuits and Systems, 2005, Kobe, Japan, 23-26 May 2005.

Low-Power Spatial Computing using Dynamic Threshold Devices

Paul Beckett
School of Electrical and Computer Engineering
RMIT University
Melbourne, Australia
pbeckett@rmit.edu.au

***Abstract*—Asynchronous spatial computing systems exhibit only localized communication, their overall data-flow being controlled by handshaking. It is therefore straightforward to determine when a particular part of such a system is active. We show that using thin-body double-gate fully depleted SOI transistors, the shift in threshold voltage that can be produced by modulating the back-gate bias is sufficient to reduce subthreshold leakage power by a factor of more than 10^4 in typical circuits. Using TBFDSOI devices in spatial computing architectures will allow overall power to be greatly reduced while maintaining high performance.**

I. Introduction

Because they are used to working in a regime where devices have been a scarce resource, computer architects have a tradition of trading off component area against performance. However, as transistors continue to shrink, micro-architecture is entering a resource-rich era, prompting an examination of alternative spatial computing models [1, 2], in which operations (and their operators) are connected in space rather than time. Spatial organizations exploit this availability of resources to expose the full parallelism available in a task - thus completing it in time proportional to the longest path through the computation rather than to the number of operations. In general, the argument is twofold: 1) when die space is no longer at a premium, heavy multiplexing of the processor datapaths is not necessary to keep the particular problem within the available silicon area [3] and 2) although they may be much larger than the minimum sized temporal design, spatial organizations may be able to achieve high computational performance [2] without the need for the hardware to "best-guess" the temporal control flow of its software (e.g. with complex features such as pre-fetch, register renaming or branch prediction).

However, it is not immediately clear whether these distributed architectures will be ultimately scalable into the nanoscale region. As device dimensions shrink, a range of potential problems will arise including reduced current drive capability (i.e. low fanout), low gain and poor reliability [4], plus more basic manufacturing issues such as the difficulty of reaching single-atom accuracy in process layers and of achieving nanometer-scale alignment between them. Regardless of these manufacturing issues, in the long term it is most likely to be simple heat dissipation considerations [5] that will limit the scaling and operating frequency of any technology based on manipulating electronic charge (such as CMOS), even at the single-electron level [6], [7]. There is an obvious tension between the desire to exploiting massive transistor counts to extract maximal parallelism and need to rein in excessive power consumption.

The two most important sources of power consumption in CMOS are dynamic switching power ($\propto aFCV^2$) and static (sub-threshold) leakage power ($\propto I_{OFF}V_{DD}$). There is a typically small contribution from short circuit current [8] plus a number of increasingly important tunneling effects – for example, through the gate oxide [9] and directly between the source and drain [10] - but these will be ignored for the purposes of this paper. As the performance of CMOS is also related to supply voltage and threshold by $F \propto (V_{DD} - V_{TH})^{\alpha}/V_{DD}$ [11], it will be decisions about supply (V_{DD}) and threshold voltage (V_{TH}) that will determine the static and dynamic power as well as the operating frequency. As supply voltage falls – thereby saving dynamic power, it will become increasingly difficult to find a *fixed* V_{TH} that optimizes both frequency and static power. In the extreme case, as V_{DD} is reduced to very low values, only sub-threshold operation may be possible.

In this paper, we examine how the idea of dynamically shifting the threshold voltage during operation can be applied to spatial computing architectures based on connected asynchronous machines. In particular, we focus on the application of thin-body (TB), fully-depleted (FD) double-gate (DG) silicon-on-insulator (SOI) devices that are likely to become a preferred nanoscale circuit element due to improved sub-threshold performance and better control of short-channel effects. FD-DGSOI devices may be ultimately scalable to gate lengths of about 10nm [12], although achieving the required level of dimensional control will be extremely difficult [13] as will achieving acceptable performance targets in the face of device parasitics.

In spatial computing, applications written in high-level languages are compiled directly into hardware circuits that exhibit only localized communication and require no global control, such as a master clock [14]. Data flow between operators is controlled by handshaking and it is therefore straightforward to determine when a particular part of the system is active. We show that large shifts in threshold voltage can be produced in nanoscale double-gate devices (in a ground-plane configuration) by modulating the back-gate voltage. By using the asynchronous interface to set the back-gate bias value for active operators, this characteristic can be exploited to substantially reduce power in spatial architectures while maintaining high performance.

The remainder of the paper proceeds as follows. Section II briefly looks at a number of previous proposals that have exploited the tradeoff between supply, threshold, performance and power in low-power systems. In Section III we present simulation results that

demonstrate the variable threshold performance of a thin-body, fully depleted double gate SOI device and demonstrate how this threshold adjustment can be used to trade off the performance of the device against its sub-threshold leakage. In Section IV, we show how these devices might be exploited in nanoscale spatial architectures and finally in Section V we conclude and point the way towards future research into this area.

II. LOW-POWER TECHNIQUES

Static power is mainly proportional to the average off current (I_{OFF}) of the transistors while active dissipation may be simply expressed as switching events (i.e. electron movements) per unit area. Unless new, disruptive cooling technologies emerge, a maximum figure of approximately $100W/cm^2$ [15] will hold regardless of the density of actual switching devices and this implies that as device density increases, both the static power per device and the device activity at a given instant must fall. Further, in portable devices a reasonable power target might be more like $0.1W/cm^2$ - reducing the power density target by at least 10^3.

As dynamic power is given by aFC_LV^2, one way of reducing it is to limit the number of devices that switch per cycle (i.e. reduce aF) by using asynchronous circuits [16]. Asynchronous circuits do not incur the significant power cost of the global clock wire that can consume more than half of the power budget in a high performance computing system. Although they have already been shown to be very power effective in allowing effective clock frequencies to fall by exploiting parallelism [14], the approach is still sensitive to the effects of static power loss.

Scaling of V_{DD} reduces dynamic power consumption but at the cost of reduced performance. This performance loss can be offset somewhat by lowering the threshold voltage V_{TH}, but at the expense of greater subthreshold leakage. Balancing the needs of low power and high performance can be a difficult optimization problem that has to be addressed at a number of levels [17], [18]. There have been many proposals to manage subthreshold power - including source biasing, exploiting the stack effect [19, 20], and static multiple threshold partitioning [21]. Multi-threshold CMOS [22] is an example of a technique that operates at the block level by gating a low V_{TH} circuit block with high V_{TH} power ("sleep") devices. In MTCMOS circuits, the sleep transistors are very large and require careful sizing to avoid problems on the virtual ground/power lines. Alternative device-level techniques have been proposed in which V_{TH} may be dynamically adjusted by applying a reverse bias to the body terminal during the standby mode or periods of low activity [23], [24]. In this way, the threshold voltage can be increased and subthreshold leakage reduced. The high load capacitance of the body terminal implies that switching between leakage modes will incur a significant dynamic power cost.

Maintaining low I_{OFF} will be increasingly difficult as channel lengths scale down below 50nm [25] and in addition, performance fluctuations due to random dopant distribution in the channel will become a major problem [26]. Out of a range of alternative structures, thin-body double-gate silicon on insulator (SOI) transistors are particularly attractive for scaling CMOS into the nanoscale regime because of their resistance to short-channel effects and the suppression of off-state leakage currents [27-29]. Random dopant effects can be avoided by using an undoped or very lightly doped channel region. Ultra-thin-body metal-gate FETs have already been demonstrated [30] and as well as double-gate devices using low-barrier silicide source and drain regions [31]. In [32], both p and n-type transistors with gate lengths down to 15nm were fabricated and it was shown that leakage currents (that have been a traditional drawback of Schottky barrier FETs) could be controlled using thin-body SOI techniques.

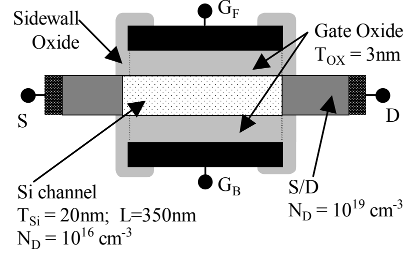


Fig. 1 Thin body fully-depleted double-gate transistor structure

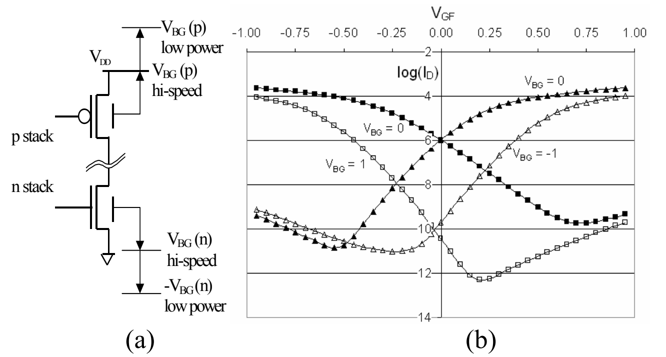
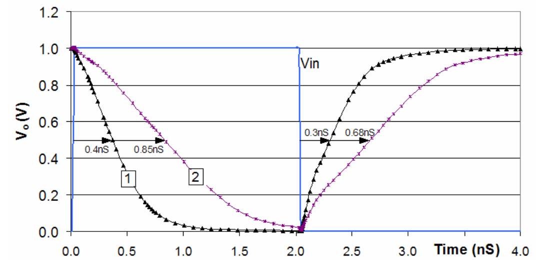


Fig. 2 (a) The general form of the transistor stack with dynamic threshold shifting. (b) Simulated TBFD-DGSOI p and n-type transistors - $\log(I_D)$ vs. front gate voltage (V_{GF}), T_{SI}=20nm. Shifting V_{BG} by ±1V effectively shifts V_{TH} by 0.3V, resulting in a reduction of I_{OFF} by a factor of ~$7x10^3$ (n-type) and ~$2x10^4$ (p-type) and an on-current reduction of ~2.2 (at $|V_{DD}|$=1).



V_{IN}	I_{OFF} (Hi-Speed)	I_{OFF} (Lo-Power)	Ratio
0	$5.58x10^{-7}$	$6.73x10^{-11}$	$8.3x10^3$
1	$1.75x10^{-6}$	$8.14x10^{-11}$	$2.15x10^4$
mean	$1.15x10^{-6}$	$7.43x10^{-11}$	$1.5x10^4$

Fig. 3 Basic inverter characteristics (FO-4): Curve 1 = high-speed mode (V_{BGP}=1V; V_{BGN}=0V); curve 2 = low-power mode (V_{BGP}=2V; V_{BGN}=-1V).

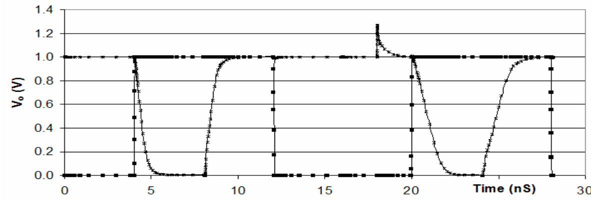
III. VARIABLE THRESHOLD IN DOUBLE-GATE DEVICES

We have simulated a number of thin-body, double-gate p and n-type devices of the general form shown in Fig. 1 using University of Florida physically based (level 10) FDSOI transistor models. Fig. 2 shows the simulated I_D vs. V_G performance for both device types with T_{SI}=20nm and T_{OX}=3nm. The channel was lightly doped at $10^{16}cm^{-3}$. For both transistors - L=0.35µm, W=1.4µm. It can be seen that by shifting the bias on the back gate (V_{BG}) by 1 volt above V_{DD} or below ground, the value of I_{OFF} (I_D for V_{GF}=0) can be reduced by a factor of more than 10^3 at the cost of a small impact on

drive current (I_D at V_{GF}=1V). This is equivalent to shifting V_{TH} by just over 0.3V. For the TBFDSOI devices studied here, $V_{BGP}=V_{DD}$ (p-type) and V_{BGN}=0 ground (n-type) sets the circuits into its high performance (and high power) mode, while $V_{BGP} = V_{DD}+1V$ and $V_{BGN} = -1V$ sets the low power mode.

The effect of this is shown in Fig. 3 for a basic inverter circuit. If we assume equal probability for the two logic states, the high performance case (curve 1) exhibits a mean static current (I_{OFF}) of 1.15µA, whereas the mean I_{OFF} for the low power case (curve 2) is $1.5x10^4$ times lower at 74pA. The leakage current for curve 2 is slightly better than the ITRS 2018 target for a single nMOS transistor in low-standby power technology (100pA/µm). Balancing the requirements of low standby power and high performance with a fixed V_{TH} will require an almost ideal value of subthreshold slope. On the other hand, uncoupling these requirements so that power and subthreshold leakage may be optimised separately will significantly relax this requirement.

In a typical realisation of the dual-gate SOI transistor, the back gate presents a load to the bias circuit that is approximately the same as that of the front gate. As illustrated in Fig. 4, switching between modes can occur at normal circuit rates with minimal disruption to the operation of the circuit. In this example using a 2-input NAND gate, the back gate biases were switched from high to low power modes at 18nS with a rise time of 500pS. The table in Fig. 4 shows the effect on the propagation delay and the subthreshold leakage for this gate. In the application envisaged here, the low-power delay times could be considered to be irrelevant as (by definition) the gates will not be operated in this mode.



Mode	T_{PHL} (nS)	T_{PLH} (nS)	Subthreshold Leakage (nA)			
			00	01	10	11
Hi-Speed	0.32	0.37	536	1090	1090	3500
Lo-Power	0.58	0.83	0.067	0.134	0.127	0.162

Fig. 4 2-NAND gate characteristics (all transistors: L=350nm, W=1.4µm). Mode switching from high speed to low power occurs at T=18nS. For clarity, only one input waveform is shown. The table shows the propagation delay times along with the subthreshold leakage for each input logic state.

IV. Exploiting Variable Threshold Devices in Asynchronous Architectures

Previous work on the application of variable threshold techniques has tended to focus on the static assignment of the high and low threshold devices within the circuit. However, as all activity in an asynchronous architecture is controlled by the handshaking signals at the interface between operators, it is an obvious strategy to use these signals to dynamically adjust the threshold according to the state of the circuits (active or inactive). Fig. 5 illustrates a standard asynchronous interface to which has been added a level shift/bias drive circuit, e.g. such a simple low-power operational amplifier circuit - that converts the ready/acknowledge signals to the V_{DD}+1 and –1V levels required for back gate biasing.

We simulated the behaviour of a simple CMOS adder circuit that is representative of the stage logic of Fig. 5. It can be seen from Table I that for this conventional full adder circuit, subthreshold current reductions in excess of $1.3x10^4$ are possible using this technique. While the actual power reduction will depend on the overall circuit activity, this tends to be very low in spatial architectures. The power overheads imposed by the back-gate bias circuit are likely to be of the same order as the adder circuit (e.g. sub-10µW), so will impact only very fine-grained organizations. This is an important issue that will be explored in future work.

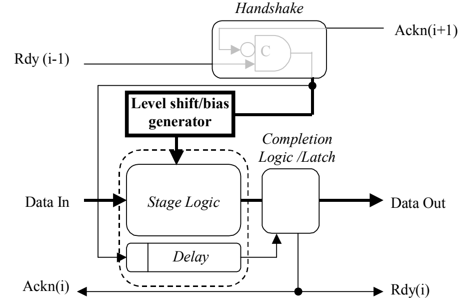


Fig. 5 Asynchronous computing stage. The only required addition to the conventional configuration is the level-shift/back gate bias generator (shown in bold).

TABLE I.
SUBTHRESHOLD LEAKAGE CURRENTS FOR A CMOS FULL-ADDER CIRCUIT (V_{DD}=1V)

A	B	C_{IN}	S	C_O	I_{OFF} (µA) (speed)	I_{OFF} (pA) (power)	Power reduction
0	0	0	0	0	6.25	477	$1.3x10^4$
0	0	1	1	0	6.89	443	$1.6x10^4$
0	1	0	1	0	6.92	446	$1.6x10^4$
0	1	1	0	1	7.50	506	$1.5x10^4$
1	0	0	1	0	7.27	473	$1.5x10^4$
1	0	1	0	1	7.55	506	$1.5x10^4$
1	1	0	0	1	7.84	487	$1.6x10^4$
1	1	1	1	1	8.75	501	$1.7x10^4$
				Mean	7.37	480	$1.54x10^4$

V. Conclusions and Future Work

We have shown how that dynamically shifting the threshold voltage during operation can reduce subthreshold power loss in double-gate transistor circuits. This will be especially applicable to organizations such as spatial computing based on connected asynchronous operators for which it is relatively straightforward to determine when a particular part of the circuit is active.

We have simulated a number of simple circuits using models for thin-body, fully-depleted, double-gate (DG) silicon-on-insulator (SOI) devices and demonstrated that subthreshold power can be reduced by a factor in excess of 10^4 using the technique. In future work, we will extend this to look at the dynamic and static power vs. performance tradeoffs that can be made in asynchronous circuits compiled directly from high level language, including the impact of the overheads imposed by the bias circuits.

References

[1] M. Budiu, "Spatial Computation," Carnegie-Mellon University, CMU CS Technical Report (Ph.D) CMU-CS-03-217, December 2003.

[2] A. DeHon, "Very Large Scale Spatial Computing," Proc. Third International Conference on Unconventional Models of Computation, UMC'02, 2002.

[3] A. DeHon, "Trends Toward Spatial Computing Architectures," IEEE International Solid-State Circuits Conference, ISSCC'99, pp. 362 - 363, 1999.

[4] W. McMahon, A. Haggag and K. Hess, "Reliability Scaling Issues for Nanoscale Devices," IEEE Transactions on Nanotechnology, vol. 2(1), pp. 33-38, 2003.

[5] S. Borkar, "Design Challenges of Technology Scaling," IEEE Micro, vol. 19(4), pp. 23-29, 1999.

[6] G. I. Bourianoff, "The Future of Nanocomputing," IEEE Computer, vol. 36(8), pp. 44-53, 2003.

[7] S. Lloyd, "Ultimate Physical Limits to Computation," Nature, vol. 406, pp. 1047 - 1054, 2000.

[8] H. J. M. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits," IEEE Journal of Solid-State Circuits, vol. 19(4), pp. 468-473, 1984.

[9] Y.-C. Yeo, T.-J. King and C. Hu, "Direct Tunneling Leakage Current and Scalability of Alternative Gate Dielectrics," Applied Physics Letters, vol. 81(11), pp. 2091 - 2093, 2002.

[10] H. Kawaura and T. Baba, "Direct Tunneling from Source to Drain in Nanometer-Scale Silicon Transistors," Japanese Journal of Applied Physics, vol. 42, Part 1(2A), pp. 351-357, 2003.

[11] K. Chen, C. Hu, P. Fang, M. R. Lin and D. L. Wollesen, "Predicting CMOS Speed with Gate Oxide and Voltage Scaling and Interconnect Loading Effects," IEEE Transactions on Electron Devices, vol. 44(11), pp. 1951-1957, 1997.

[12] Y. Kado, "The Potential of Ultrathin-Film SOI Devices for Low-Power and High-Speed Applications," IEICE Transactions on Electronics, vol. E80-C(3), pp. 443-454, 1997.

[13] Z. Ren, "Nanoscale MOSFETS: Physics, Simulation and Design," PhD Thesis, Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, 2001.

[14] M. Budiu, G. Venkataramani, T. Chelcea and S. C. Goldstein, "Spatial Computation," Proc. International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS'04, Boston, MA, 2004.

[15] SIA, International Technology Roadmap for Semiconductors - 2002 Update, Semiconductor Industry Association, 2002.

[16] I. E. Sutherland, "Micropipelines," Communications of the ACM, vol. 32(6), pp. 720 - 738, 1989.

[17] A. Basu, S.-C. Lin, V. Wason, A. Mehrotra and K. Banerjee, "Simultaneous Optimization of Supply and Threshold Voltages for Low-Power and High-Performance Circuits in the Leakage Dominant Era," Proc. Design Automation Conference, DAC 2004, San Diego, CA, pp 884-887, 2004.

[18] A. P. Chandrakasan and R. W. Brodersen, "Minimizing Power Consumption in Digital CMOS Circuits," Proceedings of the IEEE, vol. 83(4), pp. 498-523, 1995.

[19] S. Narendra, V. De, S. Borkar, D. A. Antoniadis and A. P. Chandrakasan, "Full-Chip Subthreshold Leakage Power Prediction and Reduction Techniques for Sub-0.18-mm CMOS," IEEE Journal of Solid-State Circuits, vol. 39(3), pp. 501-510, 2004.

[20] M. C. Johnson, D. Somasekhar, L.-Y. Chiou and K. Roy, "Leakage Control with Efficient Use of Transistor Stacks in Single Threshold CMOS," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 10(1), pp. 1-5, 2002.

[21] L. Wei, Z. Chen, M. Johnson, K. Roy and V. De, "Design and Optimization of Low Voltage High Performance Dual Threshold CMOS Circuits," Proc. 35th Annual Conference on Design Automation, San Francisco, California, United States, pp 489-494, 1998.

[22] J. Kao, S. Narendra and A. Chandrakasan, "Subthreshold Leakage Modeling and Reduction Techniques," Proc. 2002 IEEE/ACM International Conference on Computer-aided Design, San Jose, California, pp 141 - 148, 2002.

[23] K. Nose, M. Hirabayashi, H. Kawaguchi, S. Lee and T. Sakurai, "VTH-Hopping Scheme to Reduce Subthreshold Leakage for Low-Power Processors," IEEE Journal of Solid-State Circuits, vol. 37(3), pp. 413-419, 2002.

[24] I. Hyunsik, T. Inukai, H. Gomyo, T. Hiramoto and T. Sakurai, "VTCMOS Characteristics and is Optimum Conditions Predicted by a Compact Analytical Model," Proc. International Symposium on Low-Power Electronic Design, ISLPED'01, pp 123-128, 2001.

[25] Y.-K. Choi, D. Ha, T.-J. King and J. Bokor, "Investigation of Gate-Induced Drain Leakage (GIDL) Current in Thin Body Devices: Single-Gate Ultra-Thin Body, Symmetrical Double-Gate, and Asymmetrical Double-Gate MOSFETs," Japanese Journal Applied Physics, vol. 42, Part 1(4B), pp. 2073-2076, 2003.

[26] F. Pikus and K. Likharev, "Nanoscale Field-Effect Transistors: An Ultimate Size Analysis," Applied Physics Letters, vol. 71, pp. 3661-3663, 1997.

[27] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind and H.-S. Wong, "CMOS Scaling into the Nanometer Regime," Proceedings of the IEEE, vol. 85(4), pp. 486 - 504, 1997.

[28] C. Choi, "Modeling Of Nanoscale MOSFETS," PhD Thesis, Department of Electrical Engineering, Stanford, 2002.

[29] Q. Chen, "Scaling Limits and Opportunities of Double-Gate MOSFETs," PhD Thesis, Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, 2003.

[30] A. Vandooren, S. Egley, M. Zavala, A. Franke, A. Barr, T. White, S. Samavedam, L. Mathew, J. Schaeffer, D. Pham, J. Conner, S. Dakshina-Murthy, B.-Y. Nguyen, B. White, M. Orlowski and J. Mogab, "Ultra-Thin Body Fully-Depleted SOI Devices with Metal Gate (TaSiN) Gate, High K (HfO2) Dielectric and Elevated Source/Drain Extensions," Proc. IEEE International SOI Conference, pp 205-206, 2002.

[31] C. Hu, "SOI and Nanoscale MOSFETs," Proc. Device Research Conference, Notre Dame, IN , USA, pp 3-4, 2001.

[32] J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, T.-J. King and C. Hu, "Complementary Silicide Source/Drain Thin-Body MOSFETs for the 20 nm Gate Length Regime," Proc. International Electron Devices Meeting, IEDM2000, San Francisco, CA , USA, pp 57-60, 2000.