Ultra Low Power CMOS Technology

J. Burr and A. Peterson
Space, Telecommunications, and Radioscience Laboratory
Department of Electrical Engineering
Stanford University
Stanford, Ca. 94305
burr@mojave.stanford.edu

Abstract - **This paper discusses the motivation, opportunities, and problems associated with implementing digital logic at very low voltages, including the challenge of making use of the available real estate in 3D multichip modules, energy requirements of very large neural networks, energy optimization metrics and their impact on system design, modeling problems, circuit design constraints, possible fabrication process modifications to improve performance, and barriers to practical implementation.**

1 Introduction

As technology continues to scale into the submicron regime, massively parallel architectures are increasingly being constrained by power considerations. Minimizing the energy per operation throughout the system is assuming increasing importance. We are investigating "Ultra Low Power CMOS" to reduce the energy per operation in massively parallel signal processors, microsatellites, and large scale neural networks. We are investigating operating with supply and threshold voltages of a few hundred millivolts to reduce energy per operation by a more than 100 times.

In this paper, we show that minimum energy per operation is achieved in the subthreshold regime, and that the optimum performance is obtained when Vdd $= V_t$ and Gnd $= V_t -$ Vdd. We also show that minimum energy $\times$ time occurs when Vdd $= 3V_t$. We show that V_t should be chosen such that $I_{on}/I_{off} = ld/a$, where ld is the logic depth and a is the activity ratio, the fraction of gates which are switching at any given time. We also show that $ld = 11$ minimizes energy in a 32x32 bit parallel multiplier.

2 Motivation

The application domains we are targeting include wideband spectrometers requiring 10^{12} operations per second, microsatellites with 100mW power budgets, large scale neural networks requiring 10^{15} connections per second and 1fJ per connection, and small, massively parallel digital signal coprocessors.

As an example, a single SBus slot in a Sun SPARCstation occupies about 200cm^3, can accommodate over 2000cm^2 of active silicon using 3D stacked multichip module technology, and has a power budget of 10W (see Fig 1). An architecture with a power density of 2W/cm^2 and 40 MIPS per chip, typical of modern microprocessors, would dissipate 4KW

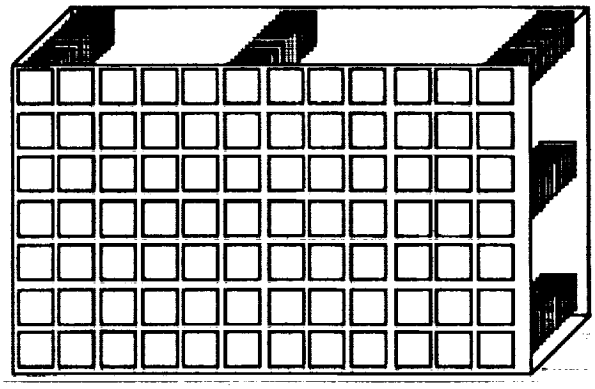

Figure 1: 3D MCM in an SBus slot: 2000 cm^2, 10W max. Vdd = 0.7V permits 10 GIPS.

if tiled over the available area and achieve 80 billion operations per second. Only 5 cm^2 of silicon can be used at 10W, yielding 200 MIPS. If the supply voltage is lowered to 700mV, each chip would dissipate 5mW, and the entire 2000cm^2 could be used to achieve 10 billion operations per second at 10W.

3 Background

Low voltage digital logic is not new. Richard Swanson described a 100mV CMOS ring oscillator in [6]. Eric Vittoz discussed subthreshold design techniques used in the digital watch industry in [4]. Carver Mead described a variety of subthreshold analog circuits for neural networks in [1]. We believe that low voltage circuits can be used effectively for massively parallel computation in power constrained environments, and that lowering the voltage in submicron technologies has the added benefit of maintaining manageable signal frequencies at the system level.

4 Transistor Current

The following equations [6,7] describe drain current as a function of gate voltage, as shown in Fig 2.

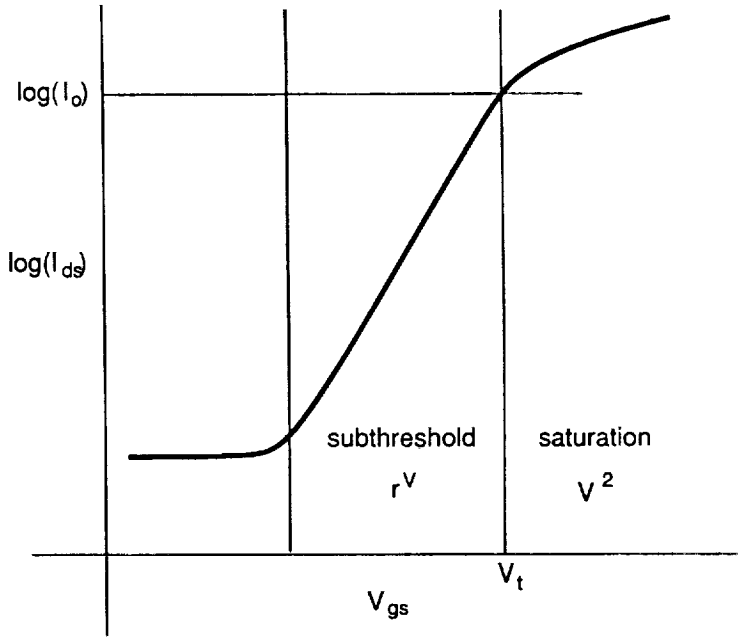

Figure 2: Transistor current vs voltage.Current in exponential with voltage below V_t, and quadratic above V_t.

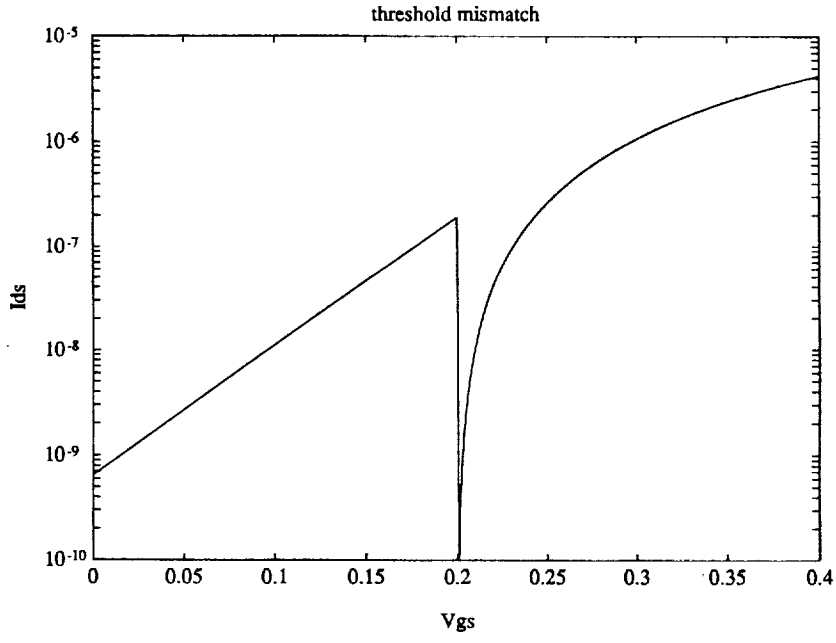

Figure 3: Model discontinuity at $V_{gs} = V_t$. The subthreshold model says $I_{ds} = knV_T^2$. The saturation model says $I_{ds} = \frac{k}{2}(V_{gs} - V_t)^2 = 0$. In the figure $V_t = 200\text{mV}$.

subthreshold: $V_{gs} < V_t$; $I_0 = knV_T^2$
$I_{ds} = I_0 e^{\frac{V_{gs}-V_t}{nV_T}}(1 - e^{-\frac{V_{ds}}{V_T}})$

saturation: $V_t < V_{gs} < V_{ds} + V_t$
$I_{ds} = \frac{k}{2}(V_{gs} - V_t)^2$

linear: $V_{ds} + V_t < V_{gs}$
$I_{ds} = \frac{k}{2}(2(V_{gs} - V_t)V_{ds} - V_{ds}^2)$

where V_{gs} is the gate-source voltage, V_t is the threshold voltage, I_{ds} is the drain current, k is the transconductance in A/V^2, n is the gate coupling coefficient, usually around 0.7, V_T is the thermal voltage, 0.026V, and I_0 is the current at $V_{gs} = V_t$.

Note the exponential dependence of current on voltage below V_t, and the quadratic dependence above V_t. These equations do a poor job of modeling behavior in the neighborhood of V_t (see Fig 3).

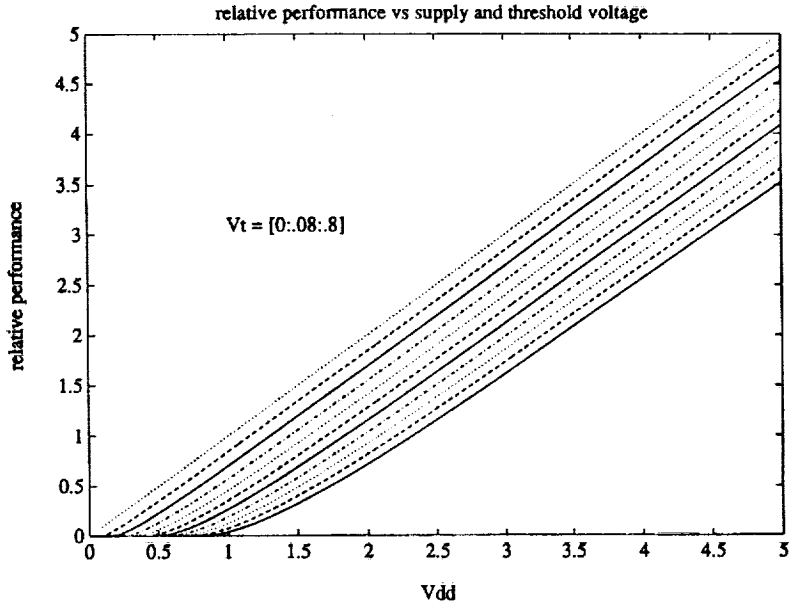

Figure 4: Performance vs voltage for different values of V_t.

Performance can be approximated when the supply voltage is over threshold by

$$f = I/Q = \frac{k}{2}(V - V_t)^2/(CV).$$

where f is the clock frequency, k is transconductance, and C is the capacitance being switched.

5 Optimum Logic Depth

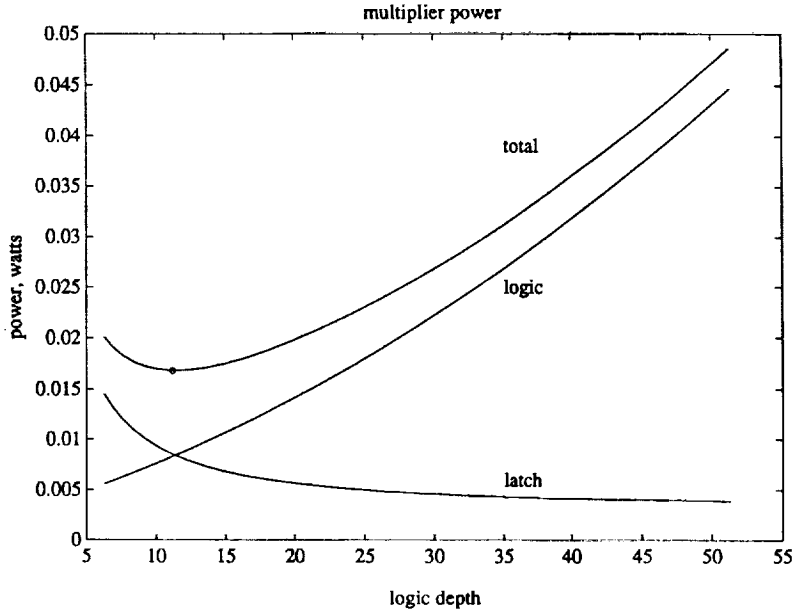

Figure 5: Optimum logic depth of a 32×32 bit tree multiplier. For a given ld, the supply voltage is lowered to match the unpiped throughput. Minimum power consumption occurs at $ld = 11$. Latch energy increases as ld decreases, eventually exceeding logic energy, which decreases as ld decreases.

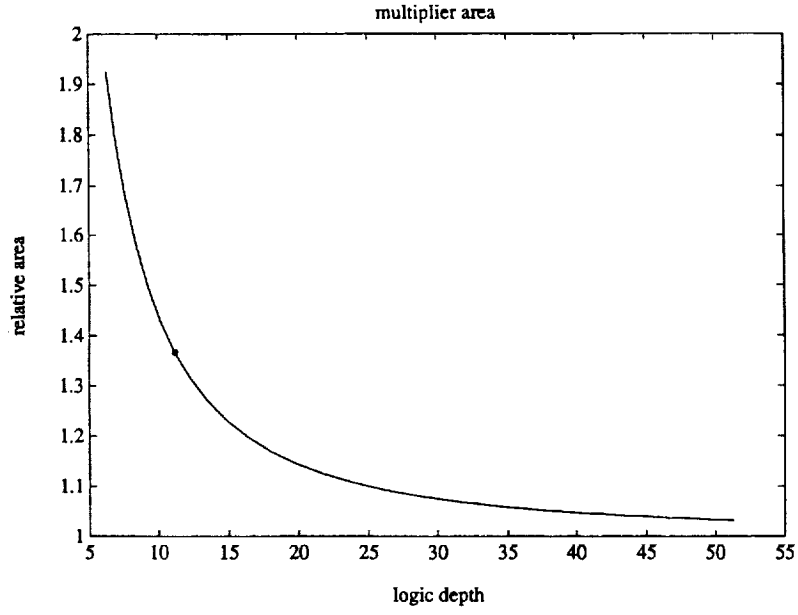

Figure 6: Relative area vs logic depth in a 32×32 bit multiplier. The area penalty at $ld = 11$ is 37%.

We found the optimum logic depth in a 32×32 bit tree multiplier by reducing the supply voltage to keep the throughput constant (see Fig 5). We also found the area penalty using this approach (see Fig. 6). $ld = 11$ is close to the propagation delay through a 4:2 adder [2].

6 Minimum Energy

The current available to switch a node is the difference between the current of the ON device and the leakage current of the OFF device. In standard CMOS, V_t is so high that I_{off} can be ignored, but in low voltage applications it can be an appreciable fraction of I_{on}:

$$\begin{aligned} t_{pd} &= \frac{Q}{I} = \frac{C_g V}{I} = \frac{C_g V}{I_{on} - I_{off}} \\ E_{dc} &= I_{off} V l_d t_{pd} = C_g V^2 \frac{l_d}{\frac{I_{on}}{I_{off}} - 1} \\ E_{ac} &= \frac{1}{2} a C_g V^2 \\ E &= E_{ac} + E_{dc} = \frac{1}{2} C_g V^2 (a + \frac{2 l_d}{\frac{I_{on}}{I_{off}} - 1}) \end{aligned}$$

E is minimum when I_{on}/I_{off} is maximum. Referring to Fig 2, I_{on}/I_{off} is maximum and constant in the subthreshold region.

In the subthreshold region, if $V_{ds} = V = V_{hi} - V_{lo}$, then $I_{on}/I_{off} = e^{(V_{hi}-V_{lo})/(nV_T)} = e^{V/(nV_T)}$, so E depends only on $V = V_{hi} - V_{lo}$. Therefore, for a given Vdd, energy is constant in the subthreshold region. For maximum performance at minimum energy, set $V_{hi} = V_t$ and $V_{lo} = V_t - V_{dd}$.

DC energy rises exponentially as Vdd decreases. AC energy rises quadratically as Vdd increases. For optimum V_t,

$$\begin{aligned} P_{ac} &= aCV^2 f \\ P_{dc} &= I_{off} V \\ I_{on} &= ldCVf \end{aligned}$$

If $P_{ac} = P_{dc}$ and Vdd = V_t, then

$$\begin{aligned} I_{on}/I_{off} &= ld/a = e^{V_t/(nV_T)} \\ V_t &= nV_T \ln(I_{on}/I_{off}) \end{aligned}$$

Figs 7 and 8 show energy vs Vdd. Table 1 lists the voltages and energies at the global minima.

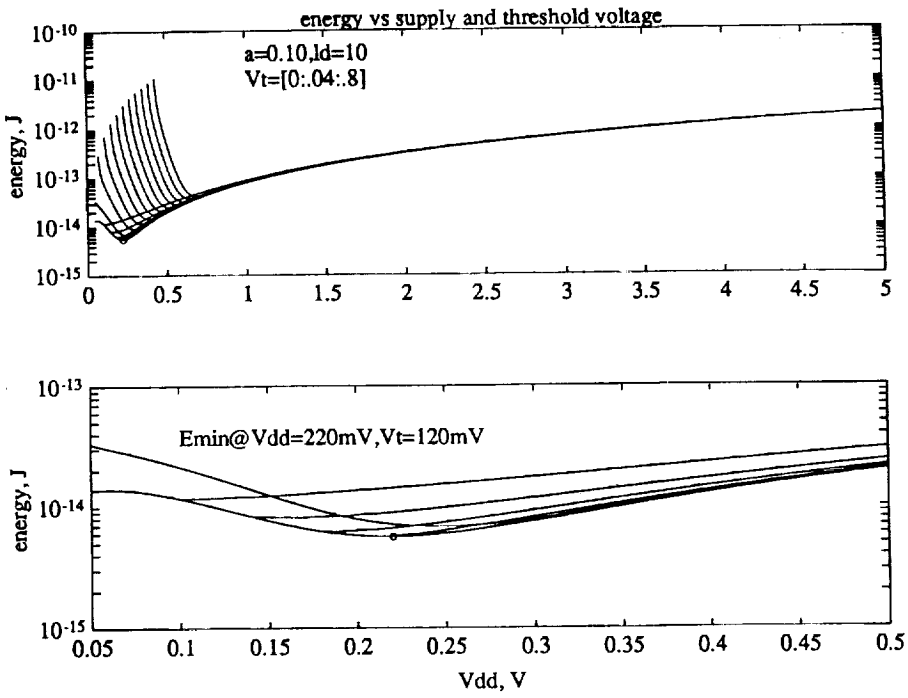

Figure 7: Energy vs supply voltage for $a = 0.10$, $ld = 10$ in 2μ CMOS

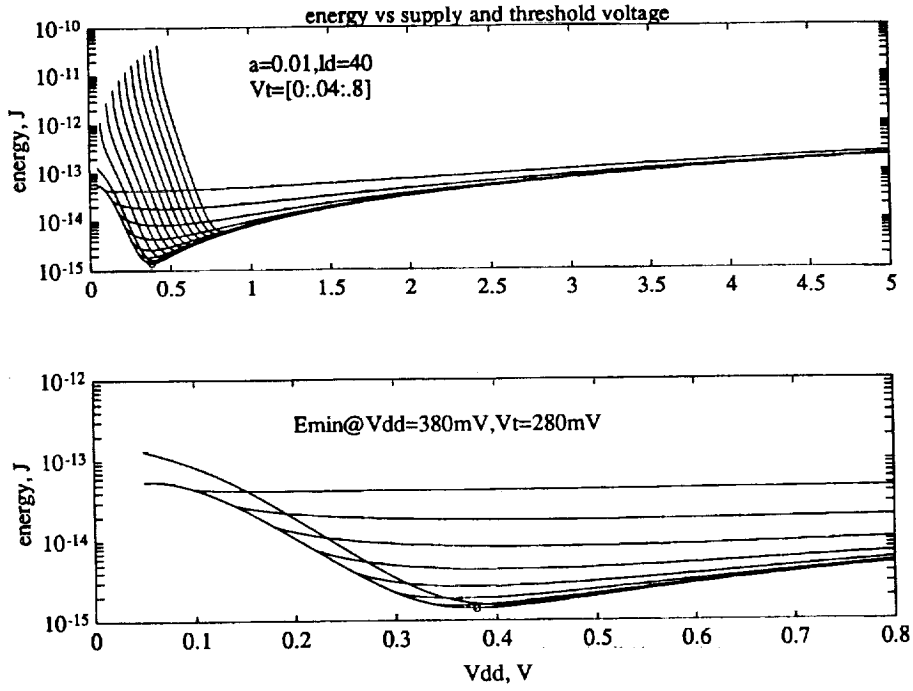

Figure 8: Energy vs supply voltage for $a = 0.01$, $ld = 40$ in 2μ CMOS

7 Minimum Energy x Time

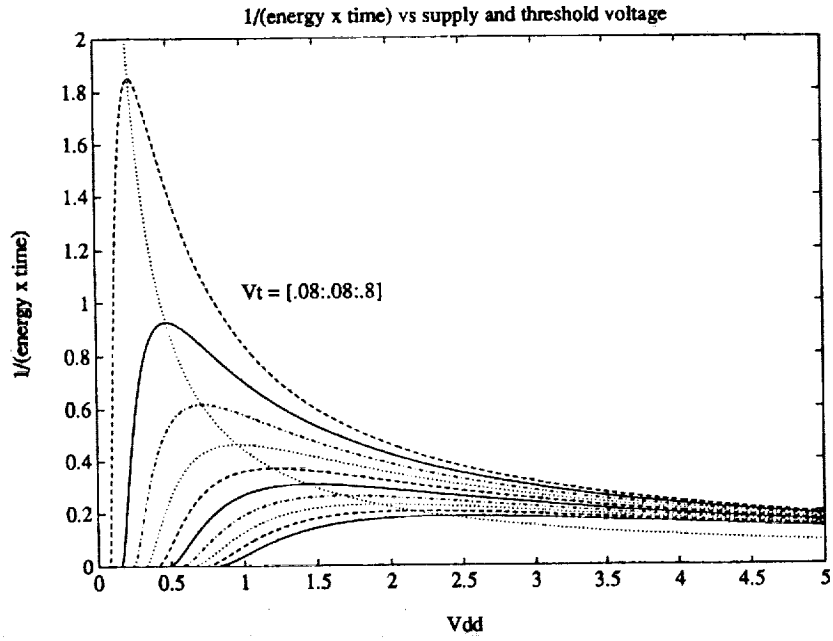

Figure 9: 1/(energy × time) vs Vdd and V_t. $Et_{\min}$ occurs at $3V_t$.

The minimum energy solution is quite slow. Performance should improve dramatically in deep submicron and with low voltage process optimizations. An alternative approach is to minimize energy × time. If we assume transistors operate mostly in saturation, then

$$\begin{aligned} Et &= V^2Q/I = V^3/(V-V_t)^2 \\ Et_{min} &= \tfrac{9}{4}V \text{ at } V = 3V_t \end{aligned}$$

Fig 9 shows a maximum at $3V_t$ which grows much more pronounced at low voltage.

8 Circuit Design Constraints

A number of interesting circuit design constraints appear when leakage currents are large, and when the dependence of current on voltage is exponential. Three constraints we have observed to date:

- Dynamic circuits are difficult to manage. A minimum size transistor will have a leakage current of about 1nA at V_t = 160mV. A dynamic storage node with 100fF of capacitance will hold 50fC of charge at Vdd=0.5V. A change of 100mV requires movement of 10fC. 10fC/1nA = 10usec.

- Exponential dependence of current on voltage makes pass transistor logic difficult to use. nfets cannot pass ones and pfets cannot pas zeros. In particular, using nfets as access transistors for static latches does not work.

parameter	negative	positive
reduce X_j	increase R_S, R_D	decrease $cjsw, cgso, cgdo$
reduce T_{ox}	decrease $V_{gs}max$ (gate-src breakdown) increase C_{ox} (increase energy)	increase k decrease n
reduce N_B	decrease $V_{ds}max$ (punchthrough)	increase uo decrease $cj, cjsw, n$
reduce N_G	increase R_G	decrease V_t
reduce N_D	increase R_S, R_D	decrease $cj, cjsw$

Table 2: Process optimization opportunities.

- Fully static logic appears to work well. Transmission gate latches work nicely. SRAM seems to work well, since one of the bitlines will be pulling down on a write.

9 Process Optimization

The opportunity exists to improve performance by optimizing fabrication processes for low voltage operation. Carrier mobility degrades significantly in submicron processes as channel doping is increased to prevent punchthrough in the presence of strong electric fields. Reduced voltage operation results in weaker fields, permitting lower channel doping which results in higher carrier mobility and increased transconductance.

Reduced voltage operation also permits lower diffusion doping, since higher diffusion resistance will not impact circuit performance due to reduced transistor drain current. This reduces diffusion capacitance to a negligible fraction of gate capacitance. The only drawback of reducing diffusion doping is that lateral diffusion is reduced, increasing the effective channel length. This is partially offset by the reduced Miller effect since the gate-drain overlap capacitance is reduced. Table 2 summarizes the impact of various process modifications on energy and performance.

While a lower bound of 60mV/decade is achievable at room temperature ($dV = nV_T ln(10)$ with $n = 1$), dV is more typically 80mV/decade in 2μ CMOS and 90mV/decade in 0.8μ CMOS. T_{ox}/d_0 can be reduced by reducing N_B, since $d_0 = \sqrt{2\epsilon_{si}\phi_{ss}/(qN_B)}$, where

$\phi_{ss} = V_T ln(N_B/n_i)$ and $n_i = \sqrt{1.5T^3 e^{-1.15/V_T}} \times 10^{16}$ [5].

Low gate, drain, and threshold voltages permit all doping concentrations to be reduced, once again due to lower electric field strength. This has two benefits for low voltage operation:

1. n is reduced, decreasing the subthreshold slope and thus reducing the supply voltage (and therefore energy per operation) necessary to achieve the desired on/off current ratio.

2. source/drain capacitances are reduced, further reducing energy per operation.

10 Barriers to Practical Implementation

A number of practical considerations place a lower bound on supply voltage. These are: external interfacing, controlling device thresholds, maintaining adequate noise margins, power supply design, power consumption of OFF devices, and circuit speed. Multichip module packaging provides the opportunity to isolate low-voltage subsystems from other system components. Limits to low voltage operation may be determined to a large extent by the power dissipation in level-shifting interface circuits. Device thresholds have been observed to vary with transistor geometry and even location on a chip [3].

A 10 watt power supply will have to deliver 20amps at Vdd = 500mV.

11 CIS Testchip

In the BiCMOS process at Stanford's Center for Integrated Systems, pfet gates are doped p+ and nfet gates are doped n+. This means that if the channel implant is excluded, both devices have thresholds close to zero volts. V_t can then be adjusted by adjusting the substrate bias voltage. We have implemented a test chip which contains a number of simple circuit structures (see Fig 10), and will hopefully have some results in time for the conference. The chip has the following characteristics:

- Pfet gates doped p+ have $V_t \approx 0V$
- Independent substrate and well biases
- self-testing convolutional coder
- ring oscillator
- VCO
- single nfet, pfet, nand, latch

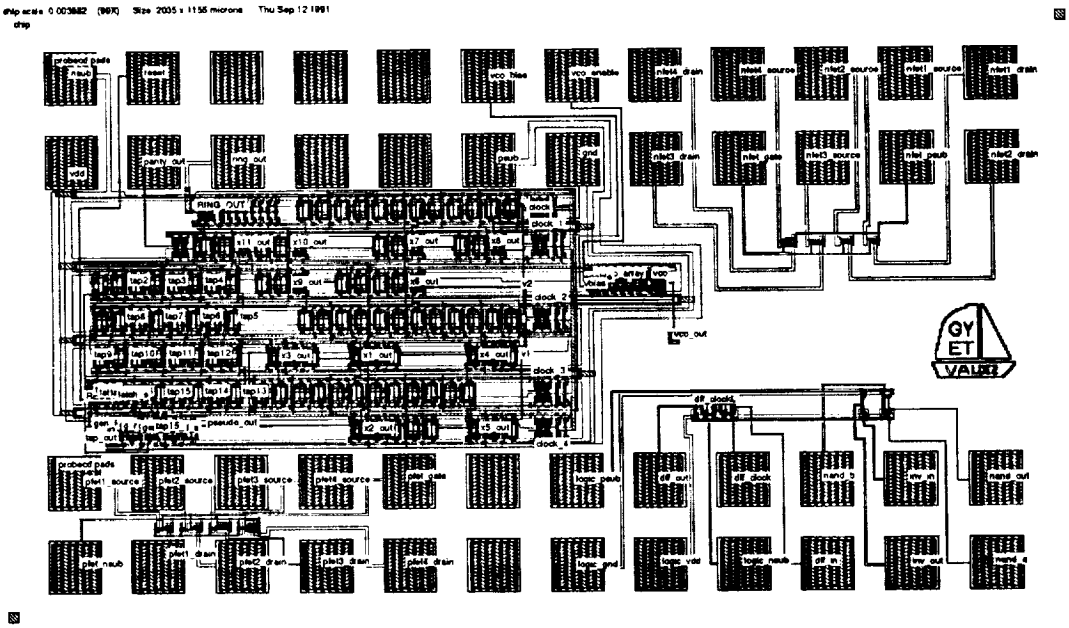

Figure 10: Ultra Low Power test chip. Separate bias voltages together with zero-V_t pfets permit threshold adjustment.

12 Conclusions

Submicron CMOS, together with 3D stacked multichip modules, and massively parallel machines demand new approaches to power dissipation. We are in the very early stages of investigating reducing energy by reducing supply and thresholds voltages. We are hopeful that low voltage CMOS can find widespread use in performance driven, power constrained systems.

13 Acknowledgements

This research was supported in part by NASA grants NAGW1910 and NAGW419, by a gift from Intel Corporation, and by a grant from Stanford's Center for Integrated Systems. Multichip modules were provided by ATT, workstations by Sun Microsystems, and VLSI fabrication by MOSIS.

References

[1] Carver A. Mead, *" Analog VLSI and Neural Systems"*, Addison-Wesley, 1989.

[2] James B. Burr and Allen M. Peterson, " Energy considerations in multichip-module based multiprocessors ", *IEEE International Symposium on Circuits and Systems*, 1991.

[3] Aleksandra Pavasovic and Andreas G. Andreou and Charles R. Westgate, " Characterization of CMOS process variations by measuring subthreshold current", *Nondestructive Characterization of Materials IV*, Plenum Press, 1991.

[4] Eric A. Vittoz, " Micropower techniques", *Design of MOS VLSI Circuits for Telecommunications*, Prentice-Hall, 1985.

[5] James R. Pfiester, " Performance limits of CMOS very large scale integration", PhD thesis, Stanford University, 1984.

[6] Richard M. Swanson, " Complementary MOS transistors in micropower circuits ", PhD thesis, Stanford University, 1974.

[7] David A. Hodges and Horace G. Jackson, *Analysis and Design of Digital Integrated Circuits*, McGraw-Hill, 1983.