Circuit Techniques for Low Power CMOS GSI

Azeez J. Bhavnagarwala, Vivek K. De, Blanca Austin and James D. Meindl
Microelectronics Research Center, School of Electrical & Computer Engineering
Georgia Institute of Technology,Atlanta, GA 30332-0269

Abstract

For a prescibed system performance, device, circuit and system design of a static CMOS datapath are conjointly optimized for different operating temperature ranges. Total power dissipation is reduced to one-third the value projected for 0.25 micron CMOS by the National Technology Roadmap for Semiconductors for a single datapath and to less than one-fourteenth the value projected for parallel datapaths assuming operation over a temperature range of 60°K above room temperature.

Introduction

We present methodologies that minimize total power drain of datapaths in static CMOS systems without loss of performance through optimal device, circuit and system design. The technology-based approach, applicable to general-purpose processors, minimizes the sum total of dynamic and static dissipation for a desired clock cycle time and a given range of operating temperatures through determination of optimal values of supply voltage and optimal values of threshold voltage and channel width for both NFET's and PFET's. In the technology-architecture based methodology, suitable for special-purpose processors running highly parallel applications, total power is further minimized for a required system throughput through determination of an optimal number of parallel datapaths in addition to the above-mentioned device and circuit parameters. In the past, approaches to minimizing dynamic power dissipation [1,2], of balancing static and dynamic power [3] and of minimizing the energy-delay product [4] advocated straightforward reductions in supply voltage and compensation of performance loss through increased parallelism of the architecture, or larger MOSFET channel width [1,2]or smaller threshold voltages[3]. Approaches to minimizing energy and the switching energy-delay product regardless of performance hâve also been reported [3,4]. The techniques presented in this paper offer the following advantages over previous approaches: (i) total power is minimized for a prescribed performance, defined by a clock frequency or a system throughput, (ii) key components of power dissipation, including both static and dynamic are considered, (iii) optimizations are performed conjointly across device, circuit and system design for both general-purpose and highly parallel special-purpose applications, (iv) a new more accurate transregional MOSFET model [5] is used, (v) the optimal design meets the performance specifications over a given operating temperature range and (vi) analytical formulations of trade-offs among device, circuit and system parameters are provided

Transregional MOSFET Models

Our transregional, analytical, physical models accurately describe MOSFET behavior in the subthreshold, saturation and linear regions of operation including high field transport effects. These new models provide smooth current-voltage characteristics across all regional boundaries to enable accurate calculation of propagation delay and total power dissipation per gate. The transregional drain current model is compared with SPICE simulations of a 0.25μ CMOS device technology in Figures 1 and 2. The competing effects of lower threshold voltage and lower carrier mobility at higher temperatures [6] on the drain current are also modeled. The performance of a datapath is then calculated using the transregional MOSFET model and system critical path models [7] that are based upon a new rigorously derived complete stochastic wiring distribution [8].

Optimal Circuit Design

The total power dissipation per critical-path gate for a prescribed cycle time,T_d, is reduced by scaling down the supply voltage, V_{dd}, while simultaneously increasing the channel width, W, and lowering the device threshold voltage, V_{to}, to compensate for the performance loss. A uniform increase in W of the critical path gates permits the performance of the datapath to improve asymptotically until the wiring capacitive load at the output of each critical path gate, determined from stochastic wiring distributions, is dominated by the gate input capacitance of the next stage. Lowering the threshold voltage permits an increase in the current drive. In both cases, that is larger W and smaller V_{to}, a constant cycle time requires a smaller supply voltage. Minimization of total power occurs when (i) the increased dynamic power drain due to larger W exceeds the corresponding reduction dynamic power due to lower V_{dd} and (ii) the exponentially increasing static dissipation due to V_{to} reduction becomes comparable to the decreasing dynamic power. The dependence of total power on cycle time and supply voltage is shown in Figure 3 The extent to which V_{dd} is scaled is limited by the above two conditions and an additional requirement of meeting the cycle time requirements over an operating range of temperature. As temperature rises, two competing effects of threshold voltage reduction and carrier mobility degradation determine the worst case performance and consequently

ISLPED 1996 Monterey CA USA
0-7803-3571-8/96/$5.00©1996

the minimum supply voltage necessary to maintain cycle time requirements over an entire range of operating temperatures. Large cycle times require low supply to threshold voltage ratios where performance improves with temperature rise as reduction in the threshold voltage with temperature dominates reductions in carrier mobility Smaller cycle times imply larger supply to threshold voltage ratios and consequently, reductions in the threshold voltage due to temperature rise do not affect the performance as much as the reduction in carrier mobility does, causing performance to degrade with temperature. Figure 4 shows these dependencies of propagation delay through the datapath on temperature. Figure 5 shows the minimum supply voltage requirements to maintain a constant cycle time over the same range of operating temperatures. Increases in temperature raise the MOSFET off-current exponentially not only due to an explicit exponential dependence on temperature but also due to reductions in threshold voltage. This dependence of static power on temperature is seen in Figure 6. Minimizing total power where the static component is significant contributes to significant fluctuations in total power with temperature. Total power is therefore minimized at the highest temperature in the operating range. The gates in the critical path are assumed to have an average fan-in of 2 and an average fan-out of 3. The delay through the 2 series NFETs of the critical path gate equals the delay through the PFET. The channel width and threshold voltage for the PFETs are chosen to minimize total power during the switching of a PFET for a given supply voltage and cycle time requirement. The simultaneous determination of V_{to}, W/L and V_{dd} at room temperature can be seen for 0.25μ CMOS in Table-I and in Figures 7 and 8, where each surface corresponds to a constant cycle time of 2.22ns or a clock frequency of 450 MHz - the 1998 performance projected by the National Technology Roadmap for Semiconductors (NTRS) [9] for high-performance 2.5V, 0.25μ CMOS processors. The bullets on these plots correspond to V_{dd} and V_{to} projected by the NTRS. Our optimal design choices of of V_{to}, W/L and V_{dd} are indicated by arrows in Figures 7 & 8

Optimal number of datapaths

Adding datapaths in parallel for a constant overall throughput, F_o, defined here as the total number of logic transitions occurring in all datapaths per unit time, permits a larger cycle time for each datapath and consequently a lower power drain from each of the datapaths due to a smaller supply voltage, a larger threshold voltage and a smaller channel width for each datapath gate. However, increasing the number of parallel datapaths, N_{paths} increases the overhead circuitry and it's dynamic power component (Figure 9) [2]. Consequently, the total power drain by all datapaths is minimized when the sum of the static and dynamic components of dissipation of the datapaths become comparable to the dynamic power drain in the overhead circuitry that must operate at the original prescribed clock frequency. Increase in the capacitance of overhead circuitry with each additional datapath is calculated from [2] where data was obtained from actual layouts.. Thus optimal choices of N_{paths}, W/L and V_{to} cumulatively and simultaneously contribute to the maximum possible reduction in V_{dd} for a constant system throughput. Fig 10 shows the dependence of total power per gate as a function of throughput for optimized single and multiple datapaths. Table I compares the total power dissipation by NTRS projections with cases of optimally designed parallel datapaths and a single datapath over a range of circuit activity and operating temperatures. Table I shows, over a temperature range of 60 K above room temperature, an optimized single data path dissipating less than one-third and optimized parallel datapaths one-fourteenth the power projected by NTRS for 2.5V, 0.25μ CMOS at a 450MHz throughput.

Conclusions

In conclusion, we report techniques to optimize CMOS circuit technology and architecture to minimize power drain for a prescribed performance over a range of operating temperatures based on transregional physical MOSFET models whose accuracy at low voltages is verified by SPICE simulations. Technology and architecture optimizations, assuming operation in a 60°K temperature range above room temperature, of 0.25μ CMOS cumulatively contribute to a reduction in V_{dd} to 490mV and reduce total power drain to one-fourteenth the value projected by NTRS for 2.5V, 0.25μ high-performance CMOS processors without any loss in throughput performance.

References

[1] A Chandrakasan et al., IEEE JSSC, Vol 27, No 4, February 1992

[2] A Chandrakasan et al., Proc. of the IEEE, Vol 83, No 4, April 1995

[3] J Burr, Hot Chips symposium V, August 1993

[4] M Horowitz et al., 1994 IEEE SLPE

[5] R Swanson et al., IEEE JSSC, Vol SC-7, No 2, April 1972 [6] C Park et al., IEEE IEDM 1995, pg 3.5.1

[7] Vivek De et al., 1996 GOMAC, pg 439-442 March 1996

[8] J Davis et al., IEEE ECTC, May 1996 (to be published)

[9] SIA Handbook, NTRS 1994

Figure 1: MOSFET gate characteristics. Comparison of HSPICE simulation of 0.25 micron CMOS technology with analytical transregional model

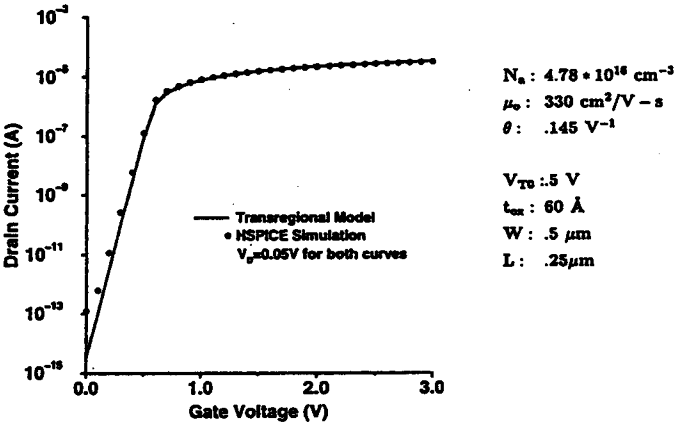


Figure 2: Propagation delay per loaded gate. Comparison of HSPICE simulations with analytical transregional model

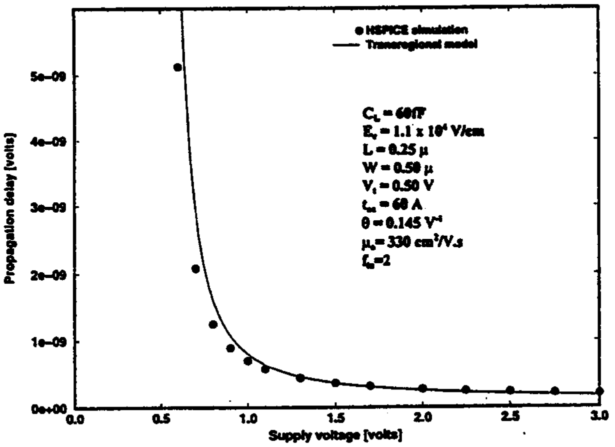


Figure 4: Dependence of propagation delay on temperature. At different supply voltages, competing effects increase or decrease the propagation delay with temperature rise

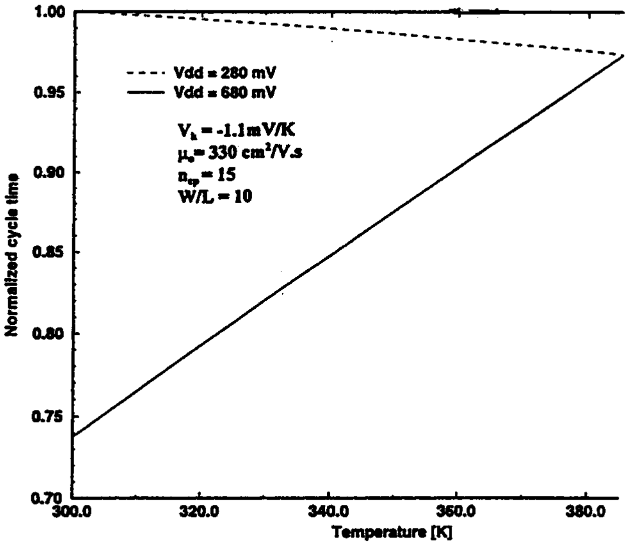


Figure 3: Dependence of total power on cycle time and supply voltage

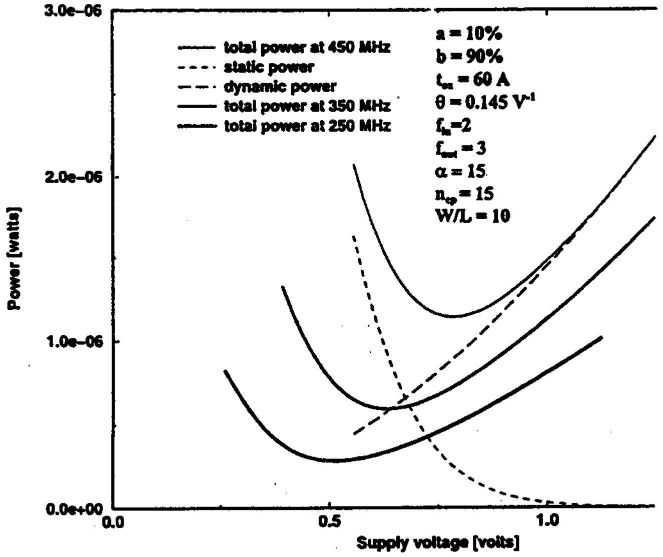


Figure 5: Minimum supply voltage required to maintain a constant cycle time over a given temperature range

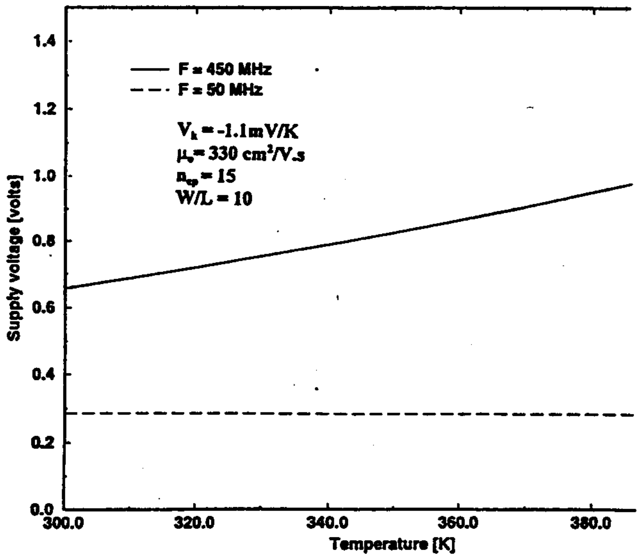


Figure 6: Exponential dependence of static power on temperature

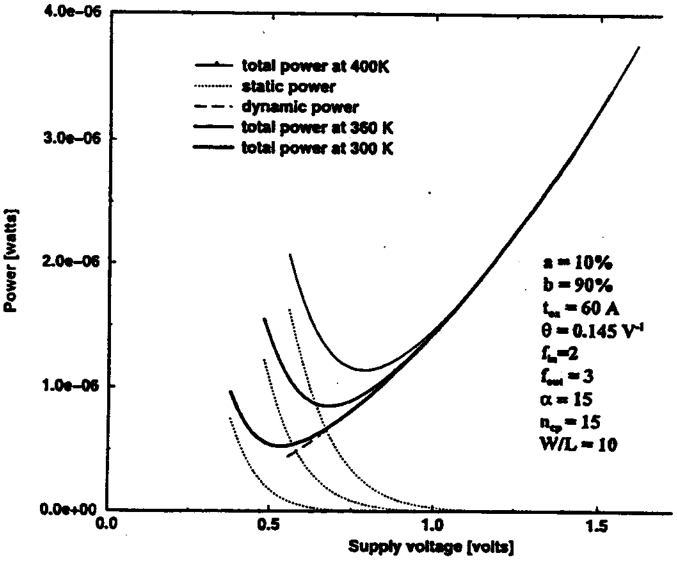


Figure 7: Channel width-to-length ratio, threshold voltage and supply voltage interdependence for a constant clock cycle time of 2.22 ns (450 MHz) for 0.25 micron CMOS

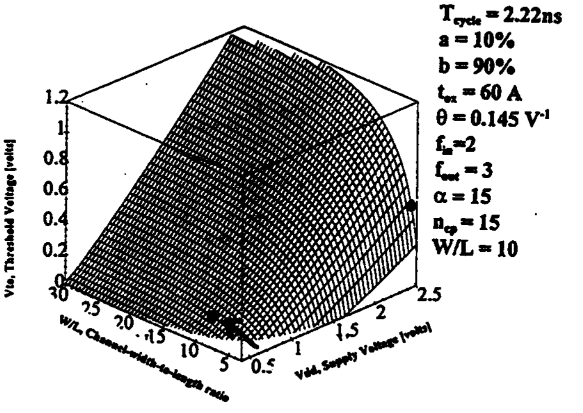


Figure 8: Total power dissipation dependence on supply voltage, channel width-to-length ratio for a constant cycle time of 2.22 ns for 0.25 micron CMOS. Threshold voltage varies along the surface to maintain cycle time

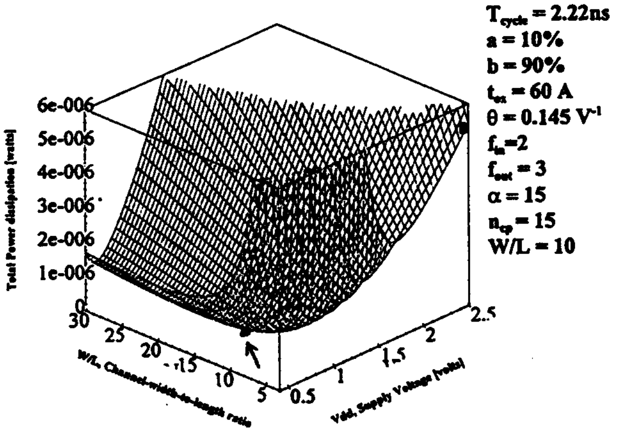


TABLE I

Single datapath, Activity = 0.1

NTRS	2500	600		3.6		5.2		
Operating temperature range ΔT	V_{dd} (mV)	V_{tn} (mV)	V_{tp} (mV)	W/L $_n$	W/L $_p$	P_{total} (μW)	P_{static} (μW)	$P_{dynamic}$ (μW)
0	753	70	30	11.3	17.5	1.08	0.14	0.94
60	950	150	90	12.1	18.7	1.75	0.26	1.47
80	1028	180	110	12.3	19.0	2.03	0.28	1.75
100	1107	210	130	12.5	19.4	2.33	0.32	2.01

Single datapath, Activity = 0.01

Operating temperature range ΔT	V_{dd} (mV)	V_{tn} (mV)	V_{tp} (mV)	W/L $_n$	W/L $_p$	P_{total} (μW)	P_{static} (μW)	$P_{dynamic}$ (μW)
0	895	140	120	10.1	12.7	0.148	0.015	0.133
60	1118	230	190	11.3	14.5	0.238	0.031	0.207
80	1218	270	210	11.6	15.4	0.276	0.030	0.246
100	1300	300	230	12.1	16.6	0.317	0.037	0.280

Parallel datapaths, Activity = 0.1

Operating temperature range ΔT	V_{dd} (mV)	V_{tn} (mV)	V_{tp} (mV)	W/L $_n$	W/L $_p$	P_{total} (μW)	Number of datapaths
0	490	110	70	3.7	4.5	0.336	4
60	510	140	110	4.1	5.3	0.352	4
80	520	150	120	4.3	5.5	0.363	4

Figure 9: Overhead circuitry and datapath components of total processor power dissipation normalized by number of gates in datapath

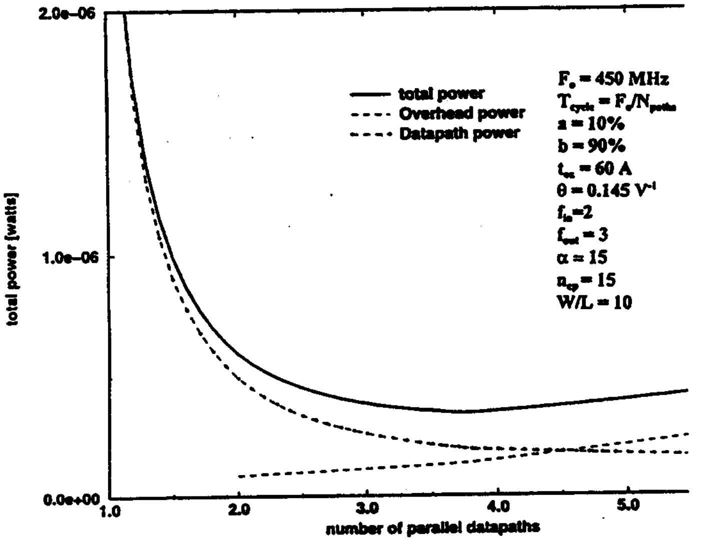
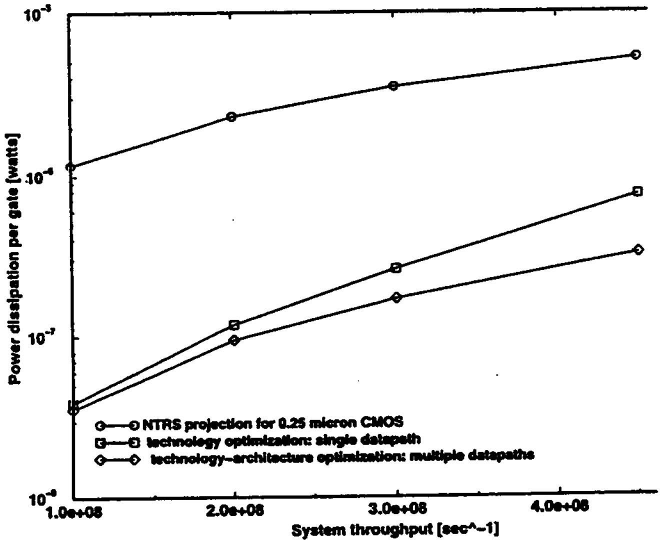


Figure 10: Comparison of NTRS projections of total power per gate with single and multiple datapath optimizations



List of Symbols used in figures

N_a : substrate doping concentration
μ_0 : low field electron mobility
θ : mobility degradation factor due to vertical fields
V_{to} : device threshold voltage
t_{ox} : gate oxide thickness
W/L : channel width-to-length ratio
C_L : total capacitive load at the output of each critical path gate
E_c : Critical field for velocity saturation
f_{in} : average fan-in of a critical path gate
f_{out} : average fan-out of a critical path gate
a : probability of switching activity
b : clock skew factor
a : ratio of interconnect to minimum feature gate input capacitance
n_{cp} : number of gates in a critical path
V_k : temperature coefficient of threshold voltage
F : clock frequency
F_0 : system throughput