

A Minimum Total Power Methodology for Projecting Limits on CMOS GSI

Azeez J. Bhavnagarwala, Blanca L. Austin, Keith A. Bowman, and James D. Meindl, *Life Fellow, IEEE*

Abstract—A circuit design methodology minimizing total power drain of a static complementary metal–oxide–semiconductor (CMOS) random logic network for a prescribed performance, operating temperature range, and short channel threshold voltage rolloff is investigated. Physical, continuous, smooth, and compact “Transregional” MOSFET drain current models that consider high-field effects in scaled devices and permit tradeoffs between saturation drive current and subthreshold leakage current are employed to model CMOS circuit performance and power dissipation at low voltages. Transregional models are used in conjunction with physical short channel MOSFET threshold voltage rolloff models and stochastic interconnect distributions to project optimal supply voltages, threshold voltages, and device channel widths minimizing total power dissipated by CMOS logic circuits for each National Technology Roadmap for Semiconductors (NTRS) technology generation. Optimum supply voltage, corresponding to minimum total power dissipation, is projected to scale to 510 mV for the 50-nm 10-GHz CMOS generation in the year 2012. Techniques exploiting datapath parallelism to further scale the supply voltage are shown to offer decreasing reductions in power dissipation with technology scaling.

Index Terms—Low-voltage CMOS, minimum power CMOS, voltage scaling.

I. INTRODUCTION

REDUCTIONS in total power dissipation of complementary metal–oxide–semiconductor (CMOS) circuit designs for ASIC’s, microprocessors, and semiconductor memories have emerged as a key design constraint over the last few years [1]. This is motivated not only by high-performance requirements in a portable environment where the size, weight and lifetime of batteries are critical, but also by heat dissipation and packaging issues in larger desktops and parallel machines as well [1]–[6]. Scaling the supply voltage for logic and memory circuits has historically been the most effective way to lower power dissipation as this reduces all components of power and is felt globally across the entire system. The 1997 National Technology Roadmap for Semiconductors (NTRS) [7] projects the supply voltage of future gigascale integrated systems to scale from 2.5 V in 1997 to 0.5 V in 2012 primarily to reduce power dissipation and power density (Fig. 1), increases of which are projected to be driven by higher clock rates, higher overall capacitance, and larger chip sizes. A key challenge in the design of logic circuits will be to meet the projected

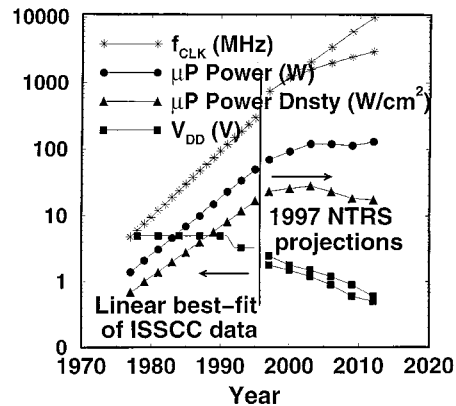


Fig. 1. Historical trends with 1997 NTRS projections for microprocessors.

performances given the competing requirements of high performance and low standby power at low voltages [1], [3], [7] in the presence of short channel threshold voltage rolloff [8].

In Section II, the physical Transregional MOSFET drain current model, verified by HSPICE simulations, is described and employed to calculate the drain current of a MOSFET in all regions of operation. In Section III, a generic CMOS datapath is modeled as a chain of critical path gates. Wiring loads at the outputs of each gate in the critical path are estimated using a stochastic interconnect distribution [9] based on Rent’s rule. These distributions have been verified for actual microprocessors. Analytical expressions for propagation delay for scaled series-connected MOSFET circuits are derived using the Transregional drain current model and verified with HSPICE simulations. Analytical expressions for the supply voltage are derived using the above model for propagation delay. Section IV introduces a simplified analysis of total power minimization. This section calculates the optimal supply voltage, minimizing total power per gate for a given performance, and elucidates the minimum power methodology implemented rigorously in Section V. A simple analytical expression for optimal supply voltage gives a rough quantitative estimate of the dependence of the optimal supply voltage on material, device, circuit, and system parameters. Section V rigorously minimizes total power drain for a single datapath using a numerical methodology. The calculations performed to determine minimum total power are consistent with the performance, technology, device count, and chip size forecasts of the NTRS. Significant threshold voltage rolloff due to short channel effects is calculated conjointly using physical models, which predict rolloff dependence on device geometry, doping profile, and supply voltage [8]. Heat removal

Manuscript received March 7, 1999; revised November 22, 1999. This work was supported by the Defense Advanced Research Project Agency under Contract F3361595C1623 and the Semiconductor Research Corporation under Contract SJ-374-002.

The authors are with the Microelectronics Research Center, Georgia Institute of Technology, Atlanta, GA 30332 USA.

Publisher Item Identifier S 1063-8210(00)04353-5.

constraints imposed (by the cost of packaging) at levels of integration and clock rates projected by the NTRS may not permit critical path gates to drive average wire lengths. High local clock frequencies are assumed to apply only within a zone of synchrony—a macrocell of a short-wire cellular array architecture whose cell size is calculated using the stochastic interconnect distribution by imposing heat removal limits on average wire length. Section VI extends the minimum power methodology to parallel datapaths by employing a simple, generic parallel datapath model to calculate the increase in overhead power dissipation with each additional parallel datapath.

II. A TRANSREGIONAL MOSFET MODEL

Compact, analytical, physical Transregional models 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 principal reason for engaging the Transregional model is that its primarily physical rather than empirical origin enables greater insight into the MOSFET parameters that are most critical to the performance of future generations of CMOS logic circuits.*

In weak inversion, where the areal inversion layer mobile carrier density is much less than the depletion region charge, the subthreshold drain current (1) is dominated by diffusion

$$I_{\text{sub}} = \frac{W}{L\eta} \mu_o C_{\text{ox}} \left(\frac{2 \frac{\eta}{\beta}}{\sqrt{1 + \frac{4\eta\theta}{\beta}} + 1} \right)^2 \cdot \exp \left[\frac{\beta}{\eta} \left(V_{gs} - V_{to} - \frac{1}{2\theta} \left(\sqrt{1 + \frac{4\eta\theta}{\beta}} - 1 \right) \right) \right] \cdot (1 - e^{-\beta V_{ds}}) \cong \frac{W}{L} \mu_o C_{\text{ox}} \frac{\eta}{\beta^2} \exp \left[\frac{\beta}{\eta} \left(V_{gs} - V_{to} - \frac{\eta}{\beta} \right) \right]. \quad (1)$$

The subthreshold slope factor $\eta = 1 + (C_d/C_{\text{ox}})$ where C_d is the channel depletion capacitance and C_{ox} the gate oxide capacitance per unit area. The device channel width-to-length ratio is given by (W/L) and μ_o is the low field carrier mobility, given by models reported in [10] for N and P channel devices. Degradation of mobility due to vertical fields in (2) and (3) is modeled with $\mu_{\text{eff}} = (\mu_o/(1 + \theta(V_{gs} - V_t)))$, where $\theta = (\mu_o/2v_{\text{norm}}t_{\text{ox}})$ [11]. The reciprocal of the thermal voltage $\beta = q/kT$.

In the linear region described by (2), the gate voltage is large enough for mobile charge density to be much greater than the depletion charge along the entire channel length and the drain current (2) is primarily determined by its drift component. In the saturation region, the gate voltage is large enough to strongly invert the channel at the source end, but the drain voltage is also large enough to cause a weak inversion region. The drain current

in the saturation region (3) is therefore given by the sum of its drift and diffusion components

$$I_{\text{linear}} = \frac{W \mu_o C_{\text{ox}}}{L(1 + \theta[V_{gs} - V_t]) \left(1 + \frac{V_{ds}}{LE_c} \right)} \cdot \left((V_{gs} - V_{to})V_{ds} - \frac{V_{ds}^2}{2} + \frac{4}{3} \phi_F \frac{Q_{BO}}{C_{\text{ox}}} \right) \cdot \left[\left(1 + \frac{V_{ds}}{2\phi_F} \right)^{3/2} - \left(1 + \frac{3}{2} \frac{V_{ds}}{2\phi_F} \right) \right]. \quad (2)$$

The gate voltage at which the device transits from the subthreshold region (weak inversion only) to the saturation region is given by (6) and (7). This voltage is determined by imposing the requirement of continuity and differentiability of the product of the field-dependent mobility and the areal mobile charge density at the boundary between these two regions

$$I_{\text{dsat}} = \frac{W}{L} (\mu_{\text{eff}}) C_{\text{ox}} \left\{ \frac{1}{\left(1 + \frac{V_{\text{dsat}}}{LE_c} \right)} \cdot \left((V_{dd} - V_{to})V_{\text{dsat}} - \frac{V_{\text{dsat}}^2}{2} + \frac{4}{3} \phi_F \frac{Q_{BO}}{C_{\text{ox}}} \right) \cdot \left[\left(1 + \frac{V_{\text{dsat}}}{2|\phi_F|} \right)^{3/2} - \left(1 + \frac{3}{2} \frac{V_{\text{dsat}}}{2|\phi_F|} \right) \right] + \frac{1}{\eta} \left(\frac{2 \frac{\eta}{\beta}}{\sqrt{1 + \frac{4\eta\theta}{\beta}} + 1} \right)^2 \{ 1 - e^{-\beta(V_{dd} - V_{\text{dsat}})} \} \right\} \quad (3)$$

$$\cong \frac{W}{L} (\mu_{\text{eff}}) C_{\text{ox}} \left\{ \frac{1}{\left(1 + \frac{V_{\text{dsat}}}{LE_c} \right)} \left((V_{dd} - V_{to})V_{\text{dsat}} - \frac{V_{\text{dsat}}^2}{2} + \frac{4}{3} \phi_F \frac{Q_{BO}}{C_{\text{ox}}} \left[\left(1 + \frac{V_{\text{dsat}}}{2|\phi_F|} \right)^{3/2} - \left(1 + \frac{3}{2} \frac{V_{\text{dsat}}}{2|\phi_F|} \right) \right] \right) \right\}. \quad (4)$$

The areal charge density of the immobile bulk charge is

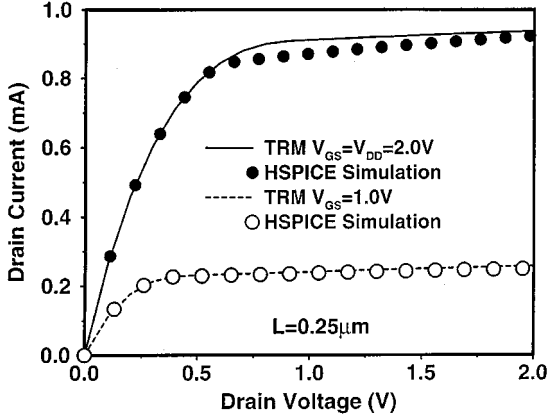
$$Q_{BO} = -\sqrt{4q\epsilon_s N_a |\phi_f|}. \quad (5)$$

The transition gate voltage from weak to strong inversion is

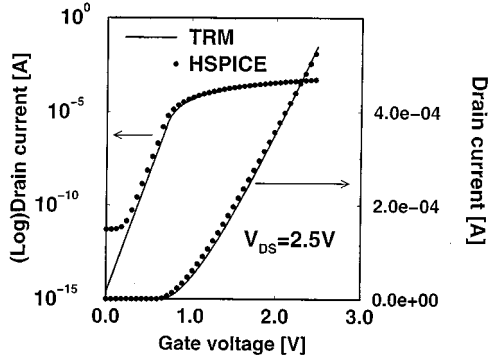
$$V_{\alpha} = V_t + \frac{\eta}{\beta} \ln \left[\frac{1}{2} \frac{\beta}{\eta} \sqrt{\frac{C_{\text{ox}}}{\eta C_d}} \left(\sqrt{1 + \frac{4\eta\theta}{\beta}} - 1 \right) \right]^2 \cong V_t + \frac{\eta}{\beta} \quad (6)$$

with

$$V_t = V_{to} + \frac{\sqrt{2q\epsilon_s N_a}}{C_{\text{ox}}} (\sqrt{V_{bs} + 2|\phi_f|} - \sqrt{2|\phi_f|}) \quad (7)$$



(a)



(b)

Fig. 2. (a) Comparison of 0.25- μm CMOS HSPICE drain characteristics with the Transregional model (TRM). $W = 2.0 \mu\text{m}$. All device parameters used are given in Appendix A. (b) Comparison of 0.25- μm CMOS HSPICE gate characteristics with the Transregional model (TRM). $W = 1.0 \mu\text{m}$. All device parameters used are given in Appendix A.

and

$$V_{to} = V_{fb} + 2|\phi_f| - \frac{\sqrt{4q\epsilon_s N_a |\phi_f|}}{C_{ox}}. \quad (8)$$

The drain saturation voltage V_{dsat} (9) is determined either by channel pinch-off or carrier velocity saturation—whichever occurs earlier at the point corresponding to $(\partial I_{linear}/\partial V_{ds})$. The critical electric field $E_c = v_{sat}/\mu_{eff}$ [12] is the lateral electric field in the device when carrier velocity reaches its saturated value v_{sat} [11]

$$V_{dsat} = LE_c \left(\sqrt{1 + \frac{2(V_{dd} - V_{to})}{LE_c \eta}} - 1 \right). \quad (9)$$

The $(1 + V_{dsat}/LE_c)$ factor in the denominator of the expressions for drain current in the linear and the saturation regions models the effects of mobility degradation due to high lateral fields. Fig. 2(a) and (b) compares the drain and gate characteristics predicted by the Transregional model with HSPICE simulations of a 0.25- μm CMOS technology. Appendix A lists the HSPICE level-3 parameters used in this simulation.

III. A SINGLE DATA-PATH MODEL

In this section, the performance of a generic CMOS processor is modeled using a simplified cycle time model, assuming a

TABLE I
AVERAGE WIRE LENGTHS AND WIRING
CAPACITANCE CALCULATIONS FROM STOCHASTIC INTERCONNECT
DISTRIBUTIONS [9] USING NTRS PROJECTIONS OF ASIC DIE SIZE AND
TRANSISTOR COUNT. NUMBERS IN ITALICS ARE OBTAINED FROM THE NTRS

Year	F (μm)	Chip size, cm^2	$N_{\text{gates}} \times 10^6$	Gate pitch GP (μm)	L_{avg} (GP)	L_{avg} (μm)	C_w (fF)
1997	0.25	4.80	6.40	8.66	8.32	72.05	43.2
1999	0.18	8.00	18.67	6.55	9.54	62.49	37.5
2001	0.15	8.5	22.67	6.12	9.78	59.85	35.9
2003	0.13	9.0	36.0	5.0	10.36	51.80	32.3
2006	0.10	10.00	66.67	3.87	11.18	43.26	25.9
2009	0.07	11.00	117.33	3.06	11.97	36.63	22.0
2012	0.05	13.00	216.67	2.45	12.90	31.61	18.9

critical path of n_{cp} two-way NAND stages, each stage driving average wire lengths and three identical gates. Average wire lengths, in units of gate pitches, are determined from stochastic interconnect distributions, derived recursively using Rent's rule, and verified for actual microprocessors [9].

The need for a progressively higher clock frequency associated with increasing average chip sizes prompted the NTRS to project "global" as well as "local" clock frequencies where the difference between the two becomes increasingly larger across the roadmap due to degradation of signal delays for long interconnects. In the analysis described in this section, critical path gates clocked at global clock frequencies drive wire lengths averaged across the entire chip, given chip sizes and transistor counts forecast by the NTRS. Critical path gates clocked at higher local frequencies drive wire lengths averaged within a macrocell of a "short-wire" cellular array architecture. The cell size is calculated using the stochastic interconnect distribution by imposing a maximum heat removal coefficient Q of 50 W/cm^2 on the average wire length of the cell.

The two-way NAND gate, with an average fan-out of three, as a basic circuit building block in the critical path, has a performance that parallels that of other circuits actually used in processor critical paths in reflecting technology improvements [13]. Propagation delay models for scaled series-connected MOSFET circuits, derived using the Transregional model, are employed to calculate the dependence of supply voltage on cycle time, logic depth, range of operating temperatures, and wiring capacitance.

A. Total Capacitance Driven by a Critical Path Gate

In logic-intensive CMOS chips, packing densities are interconnect limited [14] where the effective size of a gate is determined by its wireability [15]. The gate pitch is estimated from NTRS projections for ASIC chip size and transistor count assuming an average gate has six transistors. The gate pitch is used in calculating the average wire length in microns, for global critical paths, for each NTRS generation (Table I). Assuming equal interconnect cross-sectional dimensions and that neighboring wiring levels in a multilevel network provide an approximate ground plane, total capacitance per unit length, including fringing effects, is estimated using analytical models reported in [16].

The interconnect density function [9] predicts the number of point-to-point interconnects. Real designs, however, use wiring

TABLE II

AVERAGE WIRE LENGTHS AND WIRING CAPACITANCE IMPOSED BY HEAT REMOVAL FOR THE LAST THREE NTRS GENERATIONS. SUPPLY VOLTAGE ASSUMED EQUALS THE AVERAGE VALUE FORECAST BY THE NTRS. SIZE AND NUMBER OF MACROCELLS ARE CALCULATED USING THE STOCHASTIC WIRING DISTRIBUTION [9]. NUMBERS IN ITALICS ARE OBTAINED FROM THE NTRS. $Q = 50 \text{ W/cm}^2$, $a = 0.1$

Year	V_{dd} (V)	F (μm)	f_{clk} (local) (GHz)	Gate area A_{gate} (μm^2)	C_w (local) (fF)	L_{avg} (GP) (local)	L_{avg} (μm) (local)	$N_{\text{gates/}}$ cell $\times 10^6$	N_{cells}
2006	1.05	0.10	3.5	15.0	15.55	9.66	37.38	0.924	72
2009	0.75	0.07	6.0	9.4	11.14	8.72	26.77	0.441	266
2012	0.55	0.05	10.0	6.0	7.93	7.78	19.06	0.196	1105

“nets” that more efficiently connect the source to each of its sink terminals. Using a simple model to convert the point-to-point interconnect length to a wire net distribution, the average wire length is estimated in units of gate pitches using the expression below [9]. For the 0.25- μm generation listed in the NTRS, assuming a Rent’s exponent of $p = 0.6$ and the number of gates $N = 6.4 \times 10^6$, the wiring capacitance C_w is calculated below.

Average Wire Length (in Gate Pitches): See (10), given at the bottom of the page, where

$$\chi = 4/(f_{\text{out}} + 3). \quad (11)$$

Capacitance per unit length: $c = 2.08 \times 10^{-12}$ (F/cm) [16].

Gate Pitch (μm):

$$GP = \sqrt{A_{\text{gate}}} = \sqrt{\frac{\text{Area}_{\text{Die}}}{N_{\text{gates}}}} = 8.66. \quad (12)$$

Average Fan-Out: $f_{\text{out}} = 3$.

Thus

$$C_w = L_{\text{avg}} \times GP \times c \times f_{\text{out}} = 4.32 \times 10^{-14} \text{ F}. \quad (13)$$

Chip size, gate pitch, average interconnection length, and the wiring capacitance are listed in Table I for each of the 1997 NTRS technology generations.

With simple scaling of dimensions, power density (Q) increases with increasing integration: as capacitances scale only linearly with technology, but device count per unit area increases as the reciprocal square of feature size, energy dissipated per unit area in charging or discharging capacitances increases. The constraints imposed by the cost of packaging may thus prevent the levels of integration projected by the NTRS from achieving high local clock rates. Imposing a heat removal coefficient of 50 W/cm^2 , the maximum load capacitance driven by an average local critical path gate is calculated using

$$P_{\text{gate}} \cong \frac{1}{2} a C_L V_{dd}^2 f_{\text{clk}} \quad \text{and} \quad P_{\text{gate}} \leq Q \times A_{\text{gate}}. \quad (14)$$

For a given wiring load, the performance of a static CMOS gate increases asymptotically with increasing (W/L) ratios,

with gate delays reaching within 20% [Fig. 6(b)] of the intrinsic unloaded gate delay for

$$C_w/C_L = 0.4. \quad (15)$$

Substituting (15) into (14), we can solve for the average wiring capacitance

$$C_L \leq \frac{2QA_g}{af_{\text{clk}}V_{dd}^2} \quad \text{or} \quad C_w \leq \frac{4QA_g}{5af_{\text{clk}}V_{dd}^2}. \quad (16)$$

Substituting (16) into (13), yields the power density limited average wire length within a macrocell. This average wire length substituted in (10) yields the size and number of cells. Wiring capacitance, wire length, and cell size are tabulated in Table II above.

The total capacitance C_L driven by each gate is calculated as the sum of its three components

$$C_L = C_w + C_g + C_d \quad (17)$$

where C_g and C_d are the gate and drain capacitances, respectively.

The gate capacitance is calculated as

$$C_g = C_{go} \times f_{\text{out}} \left(\left(\frac{W}{L} \right)_n + \left(\frac{W}{L} \right)_p \right) \quad (18)$$

with C_{go} given by

$$C_{go} = L^2 C_{\text{ox}}. \quad (19)$$

The drain capacitance C_d seen at the output of an unloaded static gate has three components: 1) the gate–drain overlap capacitance for the NFET and PFET devices, C_{on} and C_{op} , respectively; 2) the drain junction bottom capacitance C_{jb} ; and 3) the junction side-wall capacitance C_{jsw} . Therefore

$$C_d = (C_{\text{on}} + C_{\text{op}}) + C_{jb} + C_{jsw}. \quad (20)$$

$$L_{\text{avg}} = \chi L_{\text{point-point}} = \chi \frac{\left(\frac{p-0.5}{p} - \sqrt{N} - \frac{p-0.5}{6(p+0.5)\sqrt{N}} + N^p \left(\frac{-p-1+4^{p-0.5}}{2(p+0.5)p(p-1)} \right) \right)}{N^{p-0.5} \frac{(-2p-1+2^{2p-1})}{2p(p-1)(2p-3)} - \frac{p-0.5}{6p\sqrt{N}} + 1 - \frac{(p-0.5)\sqrt{N}}{p-1}} = 8.32 \quad (10)$$

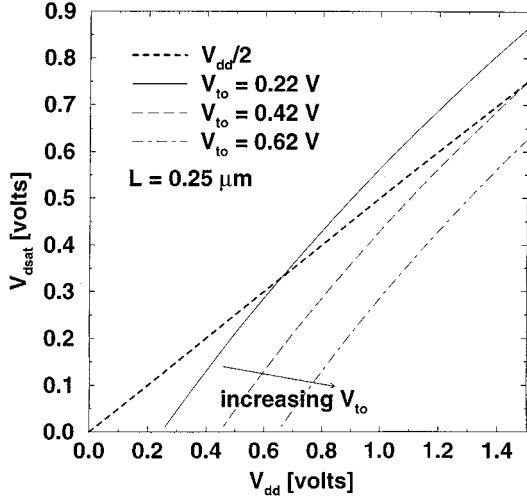


Fig. 3. V_{dsat} dependence on V_{dd} . Devices remain mostly in saturation until output node swings by $V_{dd}/2$.

From [17] we have

$$C_d = \left(\frac{1}{4} C_{go} + 3\zeta_{jb}L^2 + \zeta_{jsw}L \right) \times \left[\left(\frac{W}{L} \right)_n + f_{in} \left(\frac{W}{L} \right)_p \right] \quad (21)$$

where ζ_{jb} is the junction bottom capacitance per unit area and ζ_{jsw} is the junction side-wall capacitance per unit length. Trench isolation is assumed between active regions when calculating the sidewall capacitance. The junction bottom capacitance is calculated assuming a doping concentration that is an order of magnitude lower than that seen in the channel region [18].

B. Propagation Delay t_{pdo} of a Static CMOS Gate

Analytical expressions for propagation delay for an inverter or a generic N -input gate are derived as the time required for a 50%–50% transition [19] between the input and the output waveforms. In short channel devices where the drain saturation voltage V_{dsat} is limited by carrier velocity saturation, V_{dsat} is typically less than $\frac{1}{2}V_{dd}$. The driving transistor thus remains in the saturation region while driving its load through the first half of the transition at the output of a static CMOS gate. For the same reason, the slope of the output waveform at half the transition is, to a good approximation, independent of the slope of the input transition. Fig. 3 shows the V_{dsat} dependence on V_{dd} using (9) for threshold voltages in the neighborhood of optimal values calculated rigorously in Table III (see Section V).

The slope of the output transition at half the transition in a chain of identical symmetrical gates is thus approximated as

$$\left| \frac{dV_{out}}{dt} \right| = \frac{I_{dsat}}{C_L} \quad (22)$$

Integrating (22), we get the inverter propagation delay in response to a step

$$t_{pdo(step)} = \int_0^{V_{dd}/2} \frac{C_L dV_{out}}{I_{dsat}} = \frac{C_L V_{dd}}{2I_{dsat}} \quad (23)$$

TABLE III
PROJECTIONS UP TO YEAR 2012 CORRESPONDING TO MINIMUM POWER FOR TECHNOLOGY GENERATIONS LISTED IN THE 1997 NTRS. NUMBERS IN ITALICS ARE TAKEN FROM THE NTRS

Year	1997	2001	2006	2012
F (nm)	250	150	100	50
t_{ox} (Å)	45	27	18	8
f_{clk} (Mhz)	750	1400	2000	3000
x_f (nm)	75	45	30	15
αx_j	5%			
αL	6.7%			
V_{ddopt} (V)	1.32	1.08	0.85	0.51
V_{dsat} (V)	.74	0.59	0.46	0.27
V_{topt} (V)	.22	0.2	0.18	0.15
f_{out}	3			
n_{cp}	15			
ΔV_t (mV)	-111	-106	-80	-36
V_{ddopt}/V_{topt} (V)	5.99	5.38	4.71	3.38
$(W_n/L)_{opt}$	35	50	45	39
$(W_p/L)_{opt}$	42	59	53	45
L_{int} (GP)	8.32	9.78	11.2	12.9
C_w (fF)	43.2	35.9	25.9	18.9
C_L (fF)	174.9	147.2	92.0	50.4
Gate area ($\times 10^{-7} \text{cm}^2$)	7.5	3.75	1.5	0.6
P_{total} (μW)	19.5	25.2	14.3	4.86
P_{dyn} (μW)	11.4	11.9	6.6	1.94
P_{stat} (μW)	6.0	11.0	6.5	2.65
P_{sc} (μW)	2.1	2.3	1.2	0.27
$\frac{1}{2} C_L V^2$ (fJ)	151.7	85.8	33.2	6.5
$P_{density}$ (W/cm^2)	26.0	67.2	95.3	81
f_{clk} (max)(GHz)	0.97	2.31	4.04	7.11

The gate delay of a stage, with a finite rise time at its input, as in a chain of symmetric inverters with equal rise and fall times, is derived as the time taken for its output waveform to transit by $\frac{1}{2}V_{dd}$, less the time taken for its input waveform to move through half its complete swing

$$t_{pdo(ramp)} = t'_1 + t_2 - t_1 = 2t_1 + t_2 - t_1 = t_1 + t_2 \quad (24)$$

where $t'_1 + t_2$ is the time taken for the output waveform to transit to $\frac{1}{2}V_{dd}$ and t_1 is the time taken for the input to reach $\frac{1}{2}V_{dd}$. The delay $t'_1 (= 2t_1)$ corresponds to the time taken for the input waveform to complete its entire transition and the delay t_2 equals the time taken for the output waveform to reach $\frac{1}{2}V_{dd}$ once the input waveform has completed its transition.

From Appendix B, the closed-form expression for (24) is

$$t_{pdo(ramp)} = t_1 + t_2 = \frac{C_L V_{dd}}{2I_{dsat}} + \frac{C_L}{I_{dsat}} \cdot \left[\frac{2V_{dd} + V_T}{6} - \left(\frac{V_{dd} - V_T}{3} \right) \cdot \left(\frac{I_{dsat,33} + I_{dsat,66}}{I_{dsat}} \right) \right] \quad (25)$$

where the device threshold voltage V_{to} is modified to include reductions due to temperature and short channel effects (SCE) using

$$V_T = V_{to} + \Delta V_{to}(\Delta T) + \Delta V_{to}(SCE). \quad (26)$$

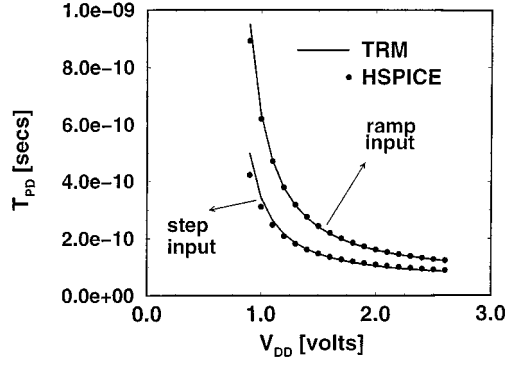


Fig. 4. Comparison of 0.25- μm CMOS HSPICE inverter propagation delay with (20) and (22). $W = 0.5 \mu\text{m}$. All device parameters are given in Appendix A.

Other terms in (25), derived in detail in Appendix B, are defined as

$$I_{\text{dsat}.33} = I_{\text{dsat}} \left(V_{\text{gs}} = \frac{V_{\text{dd}} + 2V_{\text{T}}}{3} \right) \quad (27)$$

$$I_{\text{dsat}.66} = I_{\text{dsat}} \left(V_{\text{gs}} = \frac{2V_{\text{dd}} + V_{\text{T}}}{3} \right) \quad (28)$$

$$I_{\text{dsat}} = I_{\text{dsat}}(V_{\text{gs}} = V_{\text{dd}}). \quad (29)$$

The above propagation delay models (23) and (25) are compared with HSPICE simulations in Fig. 4 for the device parameters given in Appendix A. Figs. 2–6 assume no reductions in threshold voltage due to short channel effects. Threshold rolloff using (26) is considered only in Sections V and VI.

For long channel devices, delay of a series-connected MOSFET circuit increases linearly with fan-in. This follows from a simple RC model where the resistance to the flow of current through a series-connected structure increases linearly with the number of identically sized devices. At short-channel lengths, the improved delay dependence on fan-in at short channel lengths [20], [21] brought about by velocity saturation is due to a smaller reduction in the drain saturation current with a rise in the source voltage of the topmost series-connected MOSFET. This effect on delay in a series-connected MOSFET circuit is modeled physically using (30)–(32) by calculating the fractional reduction of the normalized saturation drain current for the series-connected structure. The reciprocal of this quantity yields an “effective fan-in” f_{ineff} . Fig. 5 compares the delay predicted by this model with HSPICE and the simple RC model. Appendix A lists the level-3 HSPICE parameters used in this comparison

$$t_{\text{pd}} = t_{\text{pdo}} \times f_{\text{ineff}} \quad (30)$$

$$f_{\text{ineff}} = 1 + \frac{2(f_{\text{in}} - 1)V_{\text{dsat}} \left(1 - \frac{1}{\sqrt{2}} \right) (1 + \kappa)}{\left[(V_{\text{dd}} + V_{\text{T}}) - \frac{V_{\text{dsat}}}{2} \right]} \quad (31)$$

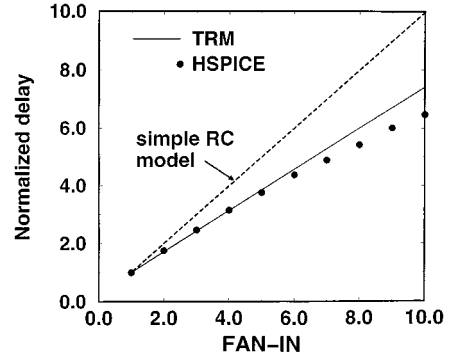


Fig. 5. Comparison of 0.25- μm CMOS HSPICE delay dependence on fan-in with (30)–(32). All device parameters are given in Appendix A.

with

$$\kappa = \frac{1}{C_{\text{ox}}} \sqrt{\frac{qN_{\text{a}}\epsilon_{\text{s}}}{4\phi_{\text{F}}}} \quad (32)$$

and f_{in} = number of series-connected devices.

The propagation delay model given by (25), when substituted into (33), yields the cycle time equation (34) that relates supply voltage (35) to device parameters, logic depth, and cycle time

$$T_{\text{cycle}} = \frac{n_{\text{cp}} t_{\text{pdo}}}{b} f_{\text{ineff}} \quad (33)$$

where b is the clock skew factor and n_{cp} the logic depth or the maximum number of gates between two clocked latches

$$T_{\text{cycle}} = \frac{1}{f_{\text{clk}}} = \frac{n_{\text{cp}} f_{\text{ineff}} C_{\text{L}}}{b I_{\text{dsat}}} \left[\frac{5V_{\text{dd}} + V_{\text{T}}}{6} - \left(\frac{V_{\text{dd}} - V_{\text{T}}}{3} \right) \left(\frac{I_{\text{dsat}.33} + I_{\text{dsat}.66}}{I_{\text{dsat}}} \right) \right]. \quad (34)$$

Solving for the supply voltage from the above cycle time equation

$$V_{\text{dd}} = \frac{\frac{2bT_{\text{cycle}}I_{\text{dsat}1}}{f_{\text{ineff}}n_{\text{cp}}C_{\text{L}}} - V_{\text{T}} \left(\frac{1}{3} + \frac{2}{3} \left(\frac{I_{\text{dsat}.33} + I_{\text{dsat}.66}}{I_{\text{dsat}1}} \right) \right)}{\frac{5}{3} - \frac{2}{3} \left(\frac{I_{\text{dsat}.33} + I_{\text{dsat}.66}}{I_{\text{dsat}1}} \right)}. \quad (35)$$

The PFET channel width that yields identical rise and fall times is calculated by equating the saturation drain currents of the two-way NAND gate that charge and discharge the load [see (36), given at the bottom of the next page].

With a rise in temperature, the two competing effects [22] of threshold voltage reduction and carrier mobility degradation determine the worst case delay and consequently the minimum supply voltage necessary to maintain cycle time requirements over an entire range of operating temperatures. Slow cycle times permit low supply to threshold voltage ratios where performance improves with temperature as reductions in the threshold voltage with temperature dominate degradation of carrier mobility. Fast cycle times translate into 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.

The temperature dependencies of threshold voltage are assumed to be [23]

$$V_{tn,p}(T+\Delta T) = V_{tn,p} + \nu_{kn,p}\Delta T. \quad (37)$$

The temperature coefficients of threshold voltage $\nu_{kn,p}$ are obtained by differentiating the threshold voltage w.r.t. temperature [23] and range from -1.2 to -0.6 mV/ $^\circ$ K for technology generations listed in the NTRS.

Stochastic interconnect models and physical Transregional drain current models are used in this section to derive a cycle time model that predicts the dependence of supply voltage on device, circuit, and system parameters. The stochastic interconnect length distribution is used to calculate the average wire length driven by a critical path gate switched at global clock frequencies. Heat removal constraints are used to calculate the average wire length driven at local clock rates with the size and number of macrocells calculated using the stochastic distribution given this requirement on average wire length. This cycle time model is used in Section V conjointly with short channel threshold voltage rolloff models and NTRS projections of cycle time, parameter tolerances, chip size, and transistor count to estimate the total power dissipated by a critical path gate.

IV. A SIMPLIFIED ANALYSIS OF POWER MINIMIZATION

A simplified methodology to minimize the total power dissipated by a critical path of static CMOS logic gates is presented in this section to provide physical insight into the calculation of optimal design parameters for a specified technology generation and performance. In essence, optimal values of supply voltage V_{dd} , long-channel threshold voltage V_{to} , NFET channel width W_n , and PFET channel width W_p are determined by scaling down V_{dd} and V_{to} while increasing transistor widths until the rate of change of static power is equal and opposite to that of dynamic power. NTRS projections [7] for feature size and gate oxide are assumed to provide a technology guideline.

Simple power dissipation and cycle time models are employed to elucidate the methodology of power minimization. In this simplified analysis, the total power dissipation of a static CMOS logic gate is assumed to be equal to the sum of its dynamic and static components

$$P_{\text{total}} = a \frac{1}{2} C_L V_{dd}^2 f_{\text{CLK}} + V_{dd} I_{\text{sub}}. \quad (38)$$

The activity factor a equals the average switching rate, i.e., the total number of logic transitions that occur in N clock cycles divided by N for large enough N [24]. For random logic networks, a is typically 10% [3]. This value is used throughout the analysis presented in this paper. The load capacitance C_L equals the sum of the wiring, gate output capacitance, and the input capacitance of the next stage as calculated in (13)–(21). Short-circuit power dissipation is typically 5–10% of the total power [4] and is neglected in this section, but is calculated numerically in Section V.

The clock frequency f_{CLK} equals the clock rate that is a consequence of imposing (39) and (40) and I_{sub} is the subthreshold leakage current given by (1). The cycle time equation (33) used in the analysis presented in this section assumes a gate propagation delay given by (23). The logic depth $n_{cp} = 15$ gates [25], effective fan-in $f_{\text{ineff}} = 2$, and clock skew factor $b = 0.9$.

The simplified minimum power methodology incorporates results from two revealing graphs, Fig. 6(a) and (b), to establish two valuable approximations that will enable a lucid interpretation of minimized total power. Fig. 6(a) depicts the normalized propagation delay versus the supply to threshold voltage ratio γ , with fixed V_{to} , W_n , and W_p to illustrate the T_{cycle} saturation at a

$$\frac{V_{dd}}{V_{to}} = \gamma \approx 6. \quad (39)$$

The optimum (V_{dd}/V_{to}) ratio is chosen at the approximate saturation point of the T_{cycle} curve, which occurs somewhat beyond a slope of negative one. Increasing V_{dd} beyond five times V_{to} yields diminishing improvements in performance. Fig. 6(b) examines the propagation delay versus the load-to-wiring capacitance ratio (C_L/C_W) with fixed V_{dd} and V_{to} and depicts T_{cycle} saturation at

$$\frac{C_L}{C_W} \approx 2.5. \quad (40)$$

Increasing the transistor widths such that (C_L/C_W) is greater than three leads to marginal performance improvements.

P -channel transistor widths are calculated such that rise–fall times are equal

$$W_n = f_{\text{ineff}} \frac{\mu_{\text{effp}}}{\mu_{\text{effn}}} W_p \quad (41)$$

where μ_{effn} and μ_{effp} are the effective carrier mobilities [11].

$$\left(\frac{W}{L}\right)_p = \frac{\left(\frac{W}{L}\right)_n \left(\frac{\mu_{on}}{(1 + \theta_n[V_{dd} - V_t]) \left(1 + \frac{V_{\text{dsatn}}}{LE_{cn}}\right)} \right) \left((V_{dd} - V_T)V_{\text{dsatn}} - \frac{V_{\text{dsatn}}^2}{2} + \varsigma_n \right)}{f_{\text{ineff}} \left(\frac{\mu_{op}}{(1 + \theta_p[V_{dd} - V_t]) \left(\frac{V_{\text{dsatp}}}{LE_{cp}}\right)} \right) \left\{ V_{dd} - V_T \right\} \left(V_{\text{dsatp}} - \frac{V_{\text{dsatp}}^2}{2} + \varsigma_p \right)} \quad (36)$$

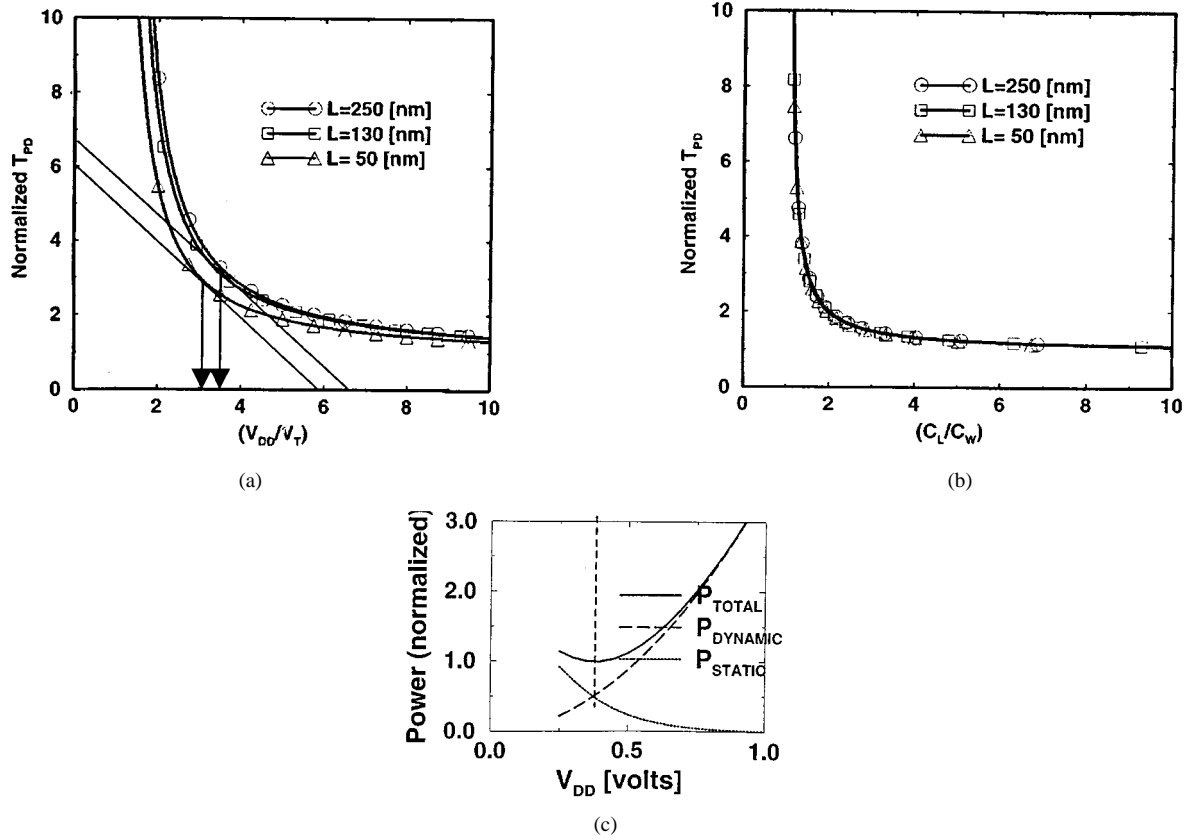


Fig. 6. (a)–(c) The simplified minimum power methodology V_{dd}/V_T and gate size are picked at points corresponding to the knee of the curves in Fig. 6(a) and (b). Optimal V_{dd} is determined by balancing static and dynamic power for the performance corresponding to this choice of V_{dd}/V_T and gate size.

Optimum values of V_{to} , W_n , and W_p are determined by imposing (39)–(41) conjointly with (43). Substituting (1) and (39) into P_{total} (38) provides P_{total} as a function of V_{dd} and γ

$$P_{total} = a \frac{1}{2} C_L V_{dd}^2 f_{clk} + V_{dd} \frac{W \mu_0 C_{ox}}{L \eta} \left(\frac{\eta}{\beta} \right)^2 \cdot \exp \left(-\frac{\beta}{\eta} \left(\frac{V_{dd}}{\gamma} + \frac{\eta}{\beta} \right) \right). \quad (42)$$

Fig. 6(c) plots the normalized P_{total} versus V_{dd} for the 0.25- μm technology generation to illustrate minimum P_{total} . The optimum V_{dd} that minimizes total power for a prescribed γ (39) is calculated by minimizing P_{total} (42) with respect to V_{dd}

$$\frac{\partial P_{total}}{\partial V_{dd}} = a C_L V_{dd} f_{clk} + \left(1 - V_{dd} \frac{\beta}{\gamma \eta} \right) \frac{W \mu_0 C_{ox}}{L \eta} \left(\frac{\eta}{\beta} \right)^2 \cdot \exp \left(-\frac{\beta}{\eta} \left(V_{dd} \gamma + \frac{\eta}{\beta} \right) \right) = 0. \quad (43)$$

Solving the previous relation for V_{dd} , yields an implicit expression for the optimum supply voltage

$$V_{ddopt} = \gamma \frac{\eta}{\beta} \left\{ \ln \left(\frac{\eta}{\beta^2} \frac{W_n}{L} \mu_0 C_{ox} \frac{1}{a C_L f_{clk}} \cdot \left(\frac{\beta}{\gamma \eta} - \frac{1}{V_{ddopt}} \right) \right) - 1 \right\}. \quad (44)$$

Assuming $(\beta/\gamma\eta) \gg (1/V_{dd})$, $V_{dsat} \approx E_C L$, and substituting (23) and (33) for f_{clk} into (44) yields the optimum supply

voltage implicitly in terms of fundamental, material, device, circuit, and system parameters

$$V_{ddopt} = \gamma \frac{\eta}{\beta} \left\{ \ln \left(\frac{1}{\beta} \frac{1}{E_C L} \frac{1 + \theta V_{ddopt} \left(1 - \frac{1}{\gamma} \right)}{\gamma - 1 - \gamma \frac{E_C L}{2 V_{ddopt}}} \frac{f_{ineff} n_{cp}}{ab} \right) - 1 \right\}. \quad (45)$$

Since V_{ddopt} is implicitly defined, an initial value for the iterative calculation of V_{ddopt} can be solved explicitly by substituting $E_C = (v_{SAT}/\mu_{eff})$ [12], assuming $\theta \ll 1$ and $(E_C L/2V_{dd}) \ll 1$, such that

$$V_{ddinitial} = \gamma \frac{\eta}{\beta} \left\{ \ln \left(\frac{1}{\beta} \frac{\mu_0}{v_{SAT} L} \frac{1}{\gamma - 1} \frac{f_{ineff} n_{cp}}{ab} \right) - 1 \right\}. \quad (46)$$

The four constraints of: 1) the (V_{dd}/V_{to}) ratio; 2) the (C_L/C_W) ratio, both corresponding to cycle time saturation; 3) equal rise and fall times; and 4) power minimization, as seen in (38)–(41), provide insights into the optimal supply voltage dependencies on system, circuit, device, and material parameters.

This simplified qualitative analysis is extended to a more complete and rigorous set of calculations in the next section and is engaged to project performance-constrained limits on CMOS energy dissipation.

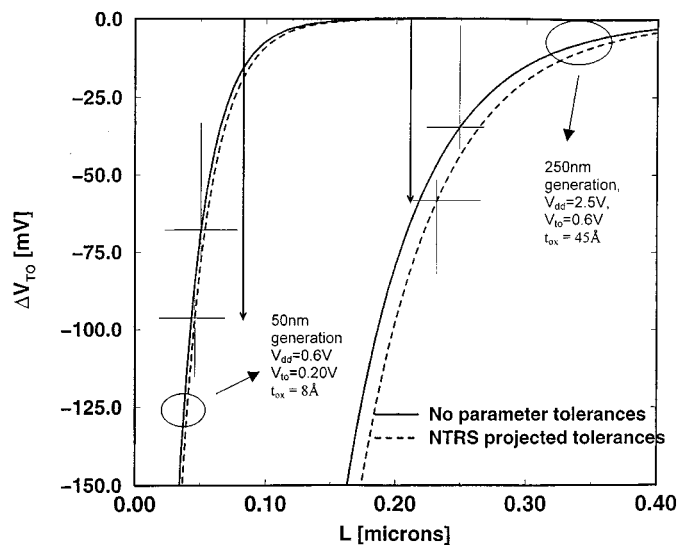


Fig. 7. Threshold voltage roll-off at NTRS-projected gate-oxide thickness, supply voltage, and long-channel threshold voltage. Dotted lines assume NTRS-projected variations of -6.7% in L . $+5\%$ variations in gate oxide thickness and source/drain junction depths are also assumed.

V. A COMPLETE MINIMUM POWER METHODOLOGY

The simplified methodology in the previous section provides physical insight into the minimum power methodology and is extended in this section to a complete numerical analysis. Several refinements are made to the simplified analysis. These encompass: 1) short channel effects of threshold reduction as well as drain-induced barrier lowering (DIBL) [8]; 2) the effects of NTRS-projected parameter tolerances on threshold roll-off; 3) temperature effects on device parameters; 4) coupling the calculation of optimal NFET and PFET channel widths simultaneously with that of the optimal supply and threshold voltages; 5) finite rise-time effects on propagation delay and total power; and 6) iterating between calculations of interdependent circuit and device parameters until convergence is reached.

Given the benefits of performance improvement, increases in level of integration and reductions of switching energy that accompany the scaling of transistor dimensions, the minimum feature size L_{\min} for a technology generation is pushed to the very edge of its physical limits [26]. The physical limit on minimum feature size is defined by the exponential threshold voltage roll-off characteristics of an MOSFET for that generation [27]. Variations in lateral and vertical device dimensions and in the supply voltage are thus bound to impact the two-dimensional (2-D) electrostatic charge coupling between the gate and source/drain regions, reducing threshold voltage and consequently increasing static power dissipation substantially [28]. Of these variations, channel length L_{\min} and gate oxide thickness t_{ox} have the most dominant effect on the reduction of threshold voltage due to the exponential dependence of threshold roll-off on these parameters [8]. The solid lines in Fig. 7 show the MOSFET roll-off characteristics calculated at NTRS-projected t_{ox} , V_{dd} , and junction depths x_j for the 50- and 250-nm generations, using physical roll-off models reported in [8]. Long channel threshold voltages necessary to yield the NTRS-projected NFET drain saturation currents of $600 \mu\text{A}/\mu\text{m}$

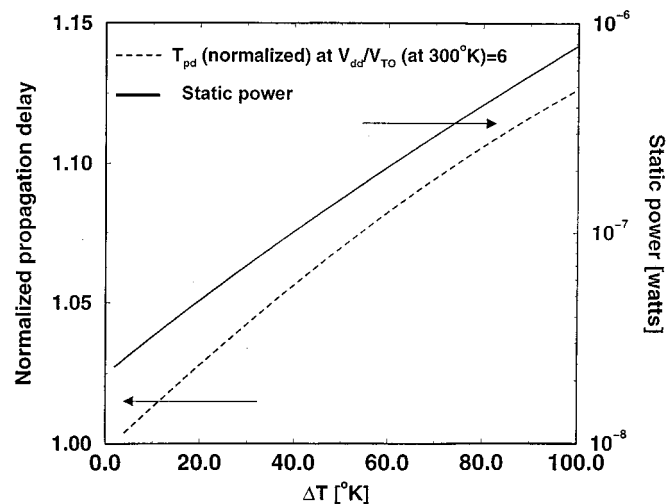


Fig. 8. Increases in cycle time and static power dissipation with temperature require the minimum power methodology to meet cycle time and minimize total power dissipation at the highest temperature in a given operating range. In this plot, $L = 250 \text{ nm}$, $\gamma = 6$, $t_{ox} = 45 \text{ \AA}$, $V_{dd} = 2.5 \text{ V}$, and $V_{to} = 0.42 \text{ V}$.

are assumed in Fig. 7. NTRS-projected variations of -6.7% in L_{\min} produce the dotted lines in Fig. 7. This deterioration also assumes a variation of $+5\%$ in t_{ox} and x_j . The roll-off curves in Fig. 7 demonstrate a total reduction in threshold voltage by 60–100 mV from long channel values, potentially increasing standby power by about an order of magnitude and thus requiring short channel threshold voltage roll-off to be considered when calculating the limits on total CMOS power.

The majority of the transistors on a chip are assumed to have minimum feature size channel lengths. The reduction in threshold voltage at this channel length due to roll-off is assumed in calculating gate delays. This calculation assumes no tolerances on device parameters. Because parameter variations of t_{ox} , L_{\min} , and x_j have a larger impact on static power than on performance, their effect on threshold voltage roll-off is considered only when calculating static power.

Between the competing effects of long channel threshold voltage reduction and mobility degradation on performance due to temperature increases, as described in Section III, mobility degradation dominates (Fig. 8) requiring the worst case delay to be met at the highest temperature for any given operating range. Static power increases exponentially with temperature, as shown in Fig. 8 due to both a decreasing long-channel threshold voltage and a lower thermal voltage, as seen in (1) and (42). Meeting delay and minimizing power at the highest temperature permit the methodology described below to guarantee that the critical path will meet cycle time for any temperature within the operating range.

The numerical calculations in the minimum power methodology sweep through a 2-D grid of long-channel threshold voltages V_{to} and NFET channel width-to-length ratios $(W/L)_n$ for a given generation as specified by its minimum feature size L_{\min} and gate oxide thickness t_{ox} . At each V_{to} on the grid, the substrate doping concentrations N_a and N_d , electron and hole saturation velocities $v_{sat,n,p}$, and the low field electron and hole mobilities $\mu_{on,p}$ are calculated using models in [10]. Models that permit calculation of the temperature and/or doping

concentration dependence [23] of bandgap energy, intrinsic carrier concentration, and conduction and valence band densities of states are used simultaneously while calculating the substrate doping concentration.

After the substrate doping concentration, carrier saturation velocities, and the low-field carrier mobilities have been calculated for a given long channel threshold voltage, the supply voltage necessary to meet the cycle time requirement, calculated using (35), is simultaneously solved for with the threshold voltage rolloff [8], the PFET channel width (36), effective fan-in (31), and total load capacitance (13)–(21). This simultaneous solution is necessary given the interdependencies of threshold voltage rolloff, PFET channel width, junction capacitance, and effective fan-in on supply voltage. The calculation iterates for a given long channel threshold voltage V_{to} and NFET channel width-to-length ratio $(W/L)_n$ until supply voltage V_{dd} , threshold rolloff ΔV_{to} , effective fan-in f_{ineff} , and load capacitance C_L converge to within a margin specified at the outset of the calculation. This calculation proceeds for a given cycle time and at a temperature of 100 K above room temperature. Fig. 9 describes the complete algorithm.

Total power dissipation is calculated as the sum of its dynamic, static, and short-circuit components for each value of V_{to} and $(W/L)_n$ and is plotted in Figs. 10 and 11, with V_{to} and $(W/L)_n$ as independent variables

$$P_{total} = P_{static} + P_{dynamic} + P_{short-circuit} \\ = I_{sub}V_{dd} + \frac{1}{2} aC_LV_{dd}^2f_{clk} + P_{sc} \quad (47)$$

where I_{sub} is given by (1) averaged over its value for NFET's and PFET's

$$P_{static} = \frac{1}{2} (I_{sub,n} + 2I_{sub,p})V_{dd}. \quad (48)$$

When calculating static power, threshold voltage rolloff is determined assuming a deviation in L_{min} of -6.7% as projected by the NTRS. A $+5\%$ increase in t_{ox} and x_j is also assumed in this calculation. The short-circuit power component during a clock cycle is calculated by numerically integrating the PFET and NFET drain currents (2) and (3) as shown in (49) and (50)

$$P_{sc,p} = aV_{dd}f_{clk} \int_0^{t_{ramp}} I_p dt \\ \text{during a rising transition at the input} \quad (49)$$

$$P_{sc,n} = aV_{dd}f_{clk} \int_0^{t_{ramp}} I_n dt \\ \text{during a falling transition at the input} \quad (50)$$

$$P_{sc} = \frac{1}{2} (P_{sc,p} + P_{sc,n}). \quad (51)$$

From Fig. 10 it can be seen that a straightforward scaling of supply and threshold voltage yields reductions in total power dissipation for a given cycle time until switching and leakage energies become comparable. Increasing the channel width independently for a given threshold voltage and cycle

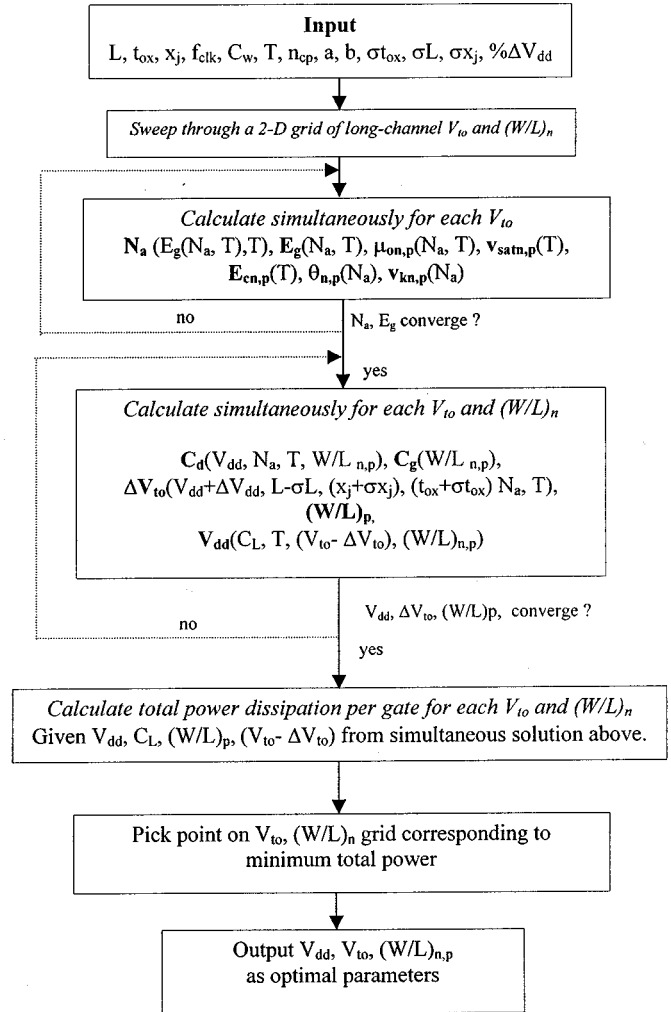


Fig. 9. Algorithm used in numerically calculating the optimal supply voltage, threshold voltage, and critical path NFET and PFET (W/L) ratios.

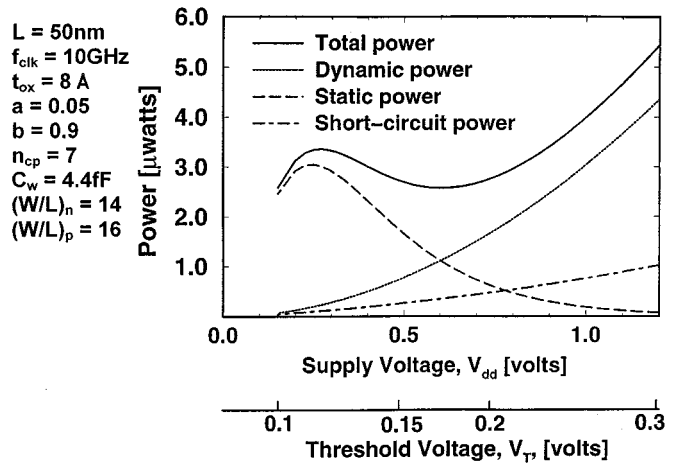


Fig. 10. Total power dissipation dependence on supply and threshold voltage.

time, as seen in Fig. 11, opens the window to further scaling of the supply voltage, until device capacitances overwhelm wiring capacitance. Beyond this point, further increases in device channel width permit only asymptotically decreasing reductions in supply voltage as shown in Fig. 11, with larger

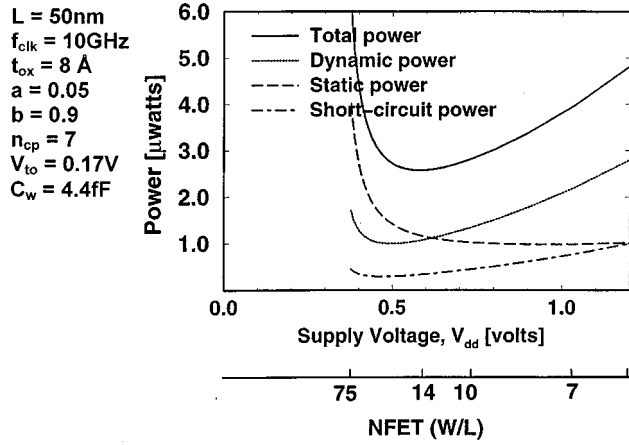


Fig. 11. Total power dissipation dependence on channel width and supply voltage.

than optimal channel widths translating into higher power dissipation due to larger gate sizes. Here we assume that if the critical path gates are small enough to be wireable, devices can be made large enough not to be dominated by wiring loads.

Thus, with threshold voltage scaling limited by static power and channel width decreases limited by wiring capacitance, the absolute minimum in total power dissipated by a CMOS gate—for a given cycle time, logic depth, percentage switching activity, and operating temperature range—corresponds to the optimal supply voltage, optimal threshold voltage, and optimal NFET and PFET channel width calculated in this analysis.

For the minimum in total power calculated above to exist, the rates of change of static and dynamic power must be equal and opposite at some point within the range of supply and threshold voltages that correspond to a given cycle time. Static power does not increase indefinitely as supply and threshold voltages are scaled and peaks at a supply voltage calculated below and shown in Fig. 12. This value of supply voltage is obtained by differentiating static power given in (42) w.r. $t.V_{dd}$ and equating this first partial derivative to zero

$$P_{\text{static}} = V_{dd} \frac{W\mu_0 C_{\text{ox}}}{L\eta} \left(\frac{\eta}{\beta}\right)^2 \exp\left(-\frac{\beta}{\eta} \left(\frac{V_{dd}}{\gamma} + \frac{\eta}{\beta}\right)\right). \quad (52)$$

$(\partial P_{\text{static}}/\partial V_{dd})$ applied to (52) yields the supply voltage corresponding to peak static power as

$$V_{dd(\text{peak}_{\text{static}})} = \frac{\eta\gamma}{\beta}. \quad (53)$$

Fig. 12 also shows that optimal supply voltages approaching this value, either due to fewer gates in the critical path or a higher activity factor, do not show a minimum in total power. Instead, total power monotonically decreases with supply voltage until the delay requirement can no longer be met. This turns out to be the case described below in Table IV for a local critical path with $n_{cp} = 7$ for the 50-nm generation operating at a local clock rate of 10 GHz. The logic depth n_{cp} is dependent on design and generation, however, this dependence is very weak and is asymptotically approaching a limit that cannot go below

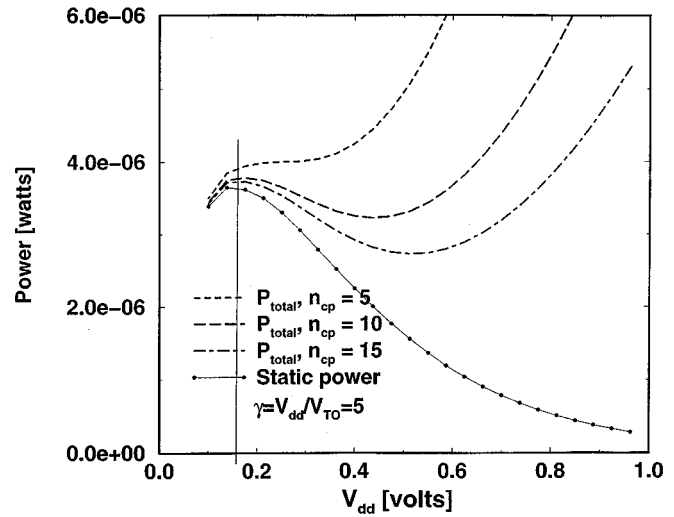


Fig. 12. Static power peaks at supply voltages equal to $\eta\gamma/\beta$. Fewer gates in the critical path and/or higher activity factors increase dynamic dissipation decreasing the optimal supply voltage until a minimum in power can no longer be seen. $L = 250$ nm, $t_{\text{ox}} = 45$ Å.

TABLE IV
PROJECTIONS UP TO YEAR 2012 CORRESPONDING TO MINIMUM POWER FOR TECHNOLOGY GENERATIONS LISTED IN THE 1997 NTRS. NUMBERS IN ITALICS ARE TAKEN FROM THE NTRS

Year	2006	2009	2012
$F(\text{nm})$	100	70	50
$t_{\text{ox}}(\text{Å})$	15	11	8
$f_{\text{clk}}(\text{Mhz})$	3500	6000	10000
$x_f(\text{nm})$	30	20	15
αx_j	5%		
σL	6.7%		
$V_{\text{ddopt}}(\text{V})$	1.05	0.75	0.55
$V_{\text{dsat}}(\text{V})$	0.44	0.32	0.21
$V_{\text{topt}}(\text{V})$	0.19	0.18	0.16
f_{out}	2		
n_{cp}	7		
$\Delta V_t(\text{mV})$	-79	-57	-39
$V_{\text{ddopt}}/V_{\text{topt}}(\gamma)$	5.52	4.17	3.44
$(W_n/L)_{\text{opt}}$	26	20	17
$(W_p/L)_{\text{opt}}$	31	22	19
$L_{\text{int}}(\text{GP})$	9.66	8.72	7.78
$C_w(\text{fF})$	15.5	11.14	7.93
$C_L(\text{fF})$	41.8	29.9	21.9
Gate area ($\times 10^{-7} \text{cm}^2$)	1.5	0.94	0.60
$P_{\text{total}}(\mu\text{W})$	16.26	10.02	6.55
$P_{\text{dyn}}(\mu\text{W})$	8.06	5.05	3.31
$P_{\text{stat}}(\mu\text{W})$	6.77	4.16	2.83
$P_{\text{sc}}(\mu\text{W})$	1.43	0.81	0.41
$\frac{1}{2} C_L V^2(\text{fJ})$	23.04	8.41	3.31
$P_{\text{density}}(\text{W/cm}^2)$	108.4	106.6	109.2

five–six. This is so because the latch delays at the start and at the end of the critical path count toward the logic depth as well and a minimum number of gates (three–four) are required for basic arithmetic and Boolean computations. The complete methodology described above is applied to several of the 1997 NTRS technology generations and optimal supply voltages, threshold

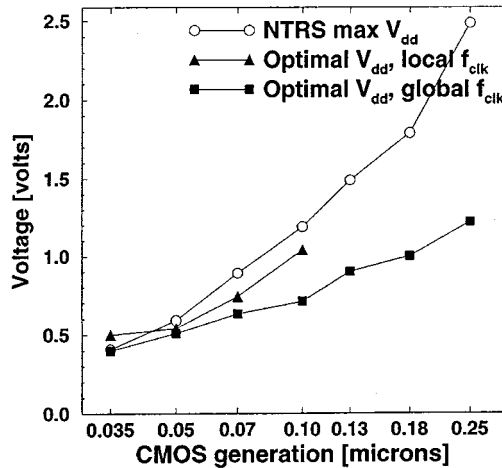


Fig. 13. Comparison of NTRS projections of V_{dd} with optimal V_{dd} calculated in Tables III and IV for local and global clock rates. n_{cp} (global) = 15 n_{cp} (local) and 7.

voltages, and NFET, and PFET channel widths calculated by the methodology are listed in Tables III and IV. The optimal V_{dd} will not permit NTRS-projected clock rates to be achieved for a conventional process with thresholds around 0.7–0.9 V. However, for processes with much lower thresholds, corresponding to optimal device threshold voltages (~ 0.3 V) gate overdrive does not deteriorate as much at lower V_{dd} , making the NTRS projections on clock rates achievable.

Average wire lengths, which track increases in chip size and transistor count, as calculated in Table I for global critical paths, or in Table II for local critical paths, are assumed driven by each critical path gate. The global critical path assumes 15 [25] two-way NAND gates with a fan-out of three. The local critical path assumes short pipeline stages with seven two-way NAND gates and a fan-out of two. Of the seven gate delays assumed in the local critical path, two correspond to gate delays due to latches.

Table III shows that while NTRS projections of global clock frequency increase by a factor of four across the roadmap, minimum feature size and gate oxide thickness are scaled more aggressively, permitting operation at lower supply voltages. From Table I, it can be seen that the average wire length, in gate pitches, increases due to larger chip sizes and higher transistor counts, permitting the average wiring capacitance to scale much less aggressively than minimum feature size. Lower optimal supply voltages accompanied by aggressive scaling of gate oxide thickness and junction depth translate into decreasing threshold rolloff permitting the optimal threshold voltage to scale to smaller values.

Tables II and IV show how rapidly average wire capacitance of a macrocell must scale to cope with a maximum heat removal rate of 50 W/cm² using (16).

Fig. 13 shows optimal supply voltage for global and local critical path gates, operating at minimum power for NTRS-projected local and global clock frequencies. Static power becomes an increasingly larger fraction of minimum total power with scaling, as seen in Tables III and IV. The impact of increasing static power can be also seen in the increase in power density for local critical path gates in Table IV.

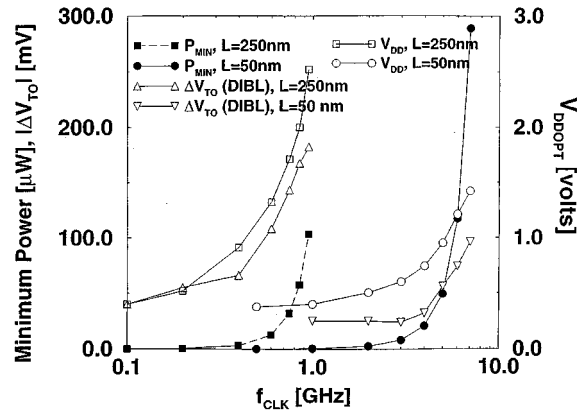


Fig. 14. Exponential increase in power at higher clock rates. Device and circuit parameters correspond to the 250- and 50-nm generations listed in Table III. $n_{cp} = 15$.

For a given technology generation, on increasing clock frequency, the optimal supply voltage increases to meet a higher performance requirement, deteriorating the threshold voltage rolloff causing static power to increase as well. Fig. 14 shows the sharp exponential rise in total minimum power as global clock rates are increased. Increasing optimal supply voltages and deteriorating threshold voltage rolloff are also plotted for the 50- and 250-nm NTRS generations in Fig. 14.

VI. PARALLEL DATAPATHS

In this section, the opportunity to reduce power drain by exploiting concurrency driven voltage scaling [2] is investigated at low supply voltages and deep submicron feature sizes. Scaling the supply voltage and compensating for the performance loss by adding datapaths in parallel so that the total number of logic operations per second (F_o) or system throughput remains constant permits the total power drain from all of the datapaths to be reduced due to a lower operating voltage [2]. However, increasing the number of parallel datapaths, N_p increases the complexity and size of the overhead circuitry required for routing, multiplexing and control of each of the parallel datapaths resulting in the dissipation of an additional component of overhead power. Also, latency increases by a factor N_p over the single datapath case. Below, a generic model for the dependence of capacitance of overhead circuitry on number of datapaths is described and used to compare power drain of an optimized single datapath with the case for parallel datapaths.

The clock requirements for each parallel datapath are reduced to

$$T_{\text{cycle}} = \frac{1}{f_{\text{clk}}} = \frac{N_p}{F_o}. \quad (54)$$

Power drain of N_p datapaths, each operating at the above clock rate given is given by

$$\begin{aligned} N_p P_{\text{datapath}} &= \frac{1}{2} a N_p C_{\text{datapath}} V_{dd}^2 \frac{F_o}{N_p} + P_{\text{static}} N_p \\ &= \frac{1}{2} a C_{\text{datapath}} V_{dd}^2 F_o + P_{\text{static}} N_p \end{aligned} \quad (55)$$

where C_{datapath} is the total switching capacitance along the critical path of a datapath and P_{static} is the total static power dissipated by each datapath. The increase in the switching capacitance of the overhead circuitry with additional datapaths is calculated relative to the datapath capacitance using the following generic model:

$$\frac{C_{\text{overhead}}}{C_{\text{datapath}}} = mN_p^\omega + \Gamma. \quad (56)$$

The parameter m models the complexity of the *control circuitry* and/or any other component of the overhead that does not increase with each additional datapath. The exponent of N_p , ω models the rate at which the *routing and multiplexing* requirements increase with each additional datapath and is specified by the size and complexity of the datapath. For an 8-b adder/comparator datapath [29], data from layouts showed an approximately quadratic dependence in (56) on N_p . Examples in [29] indicate a range on m from 0.1 to 0.7.

The dynamic power dissipated by the overhead circuitry and the datapaths is given by

$$\begin{aligned} P_{\text{overhead}}^{\text{dynamic}} &= \frac{1}{2} a C_{\text{overhead}} V_{dd}^2 F_o \\ &= \frac{1}{2} a C_{\text{datapath}} \times m (N_p^2 - 1) V_{dd}^2 F_o. \end{aligned} \quad (57)$$

Static power dissipated by the overhead circuitry is assumed to increase linearly with the overhead capacitance. The overhead circuitry capacitance is assumed to be dominated by device capacitances due to the highly local nature of their placement

$$P_{\text{overhead}}^{\text{static}} = \frac{C_{\text{overhead}}}{C_{\text{datapath}}} P_{\text{datapath}}^{\text{static}}. \quad (58)$$

Total overhead power

$$P_{\text{overhead}} = P_{\text{overhead}}^{\text{dynamic}} + P_{\text{overhead}}^{\text{static}}. \quad (59)$$

The sum total of datapath and overhead power

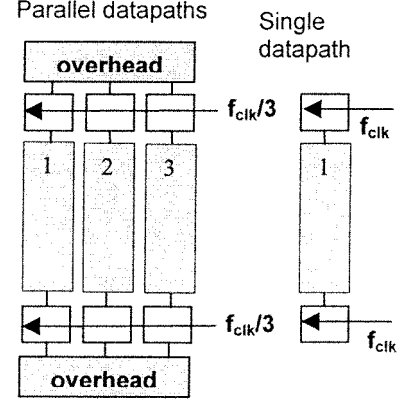
$$P_{\text{total}} = N_p P_{\text{datapath}} + P_{\text{overhead}} \quad (60)$$

or

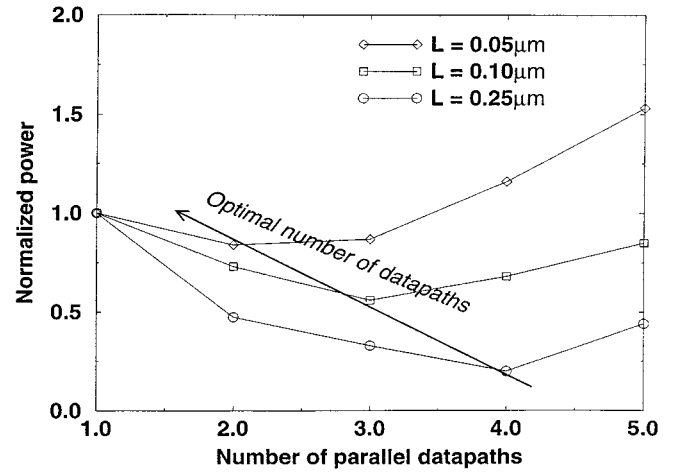
$$\begin{aligned} P_{\text{total}} &= \frac{1}{2} a C_{\text{datapath}} V_{dd}^2 F_o \left[1 + \frac{C_{\text{overhead}}}{C_{\text{datapath}}} \right] \\ &\quad + P_{\text{datapath}}^{\text{static}} \frac{C_{\text{overhead}}}{C_{\text{datapath}}} + P_{\text{datapath}}^{\text{static}} N_p. \end{aligned} \quad (61)$$

This total power dissipation is normalized by dividing by N_g , the number of gates in each datapath. Using $C_L = C_{\text{datapath}}/N_g$ where C_L is the load at the output of each datapath gate. The total power dissipation from all of the parallel datapaths, normalized by the number of gates in each datapath is given by

$$\begin{aligned} P_{\text{total}}^{\text{normalized}} &= \frac{1}{2} a C_L V_{dd}^2 F_o \left[1 + \frac{C_{\text{overhead}}}{C_{\text{datapath}}} \right] \\ &\quad + P_{\text{static}}^{\text{datapath}} \left(\frac{C_{\text{overhead}}}{C_{\text{datapath}}} + N_p \right). \end{aligned} \quad (62)$$



(a)



(b)

Fig. 15. (a) and (b) Normalized power dependence on number of parallel datapaths. Decreasing V_{dd}/V_{to} ratios across the roadmap yields smaller reductions in total power by exploiting datapath parallelism.

Fig. 15 plots the above normalized power per gate for three generations in the NTRS showing smaller reductions in total power obtainable by employing parallel datapaths.

Decreasing V_{dd}/V_{to} ratios projected by the roadmap (Table V) increases the speed penalty for a given reduction in supply voltage, increasing the required number of parallel processors significantly to compensate for the loss in performance.

VII. CONCLUSION

The limits on CMOS energy dissipation shown to be imposed by static power and by wiring capacitance are investigated using a methodology that conjointly employs physical short-channel MOSFET drain current and threshold voltage rolloff models in tandem with stochastic wiring distributions. This methodology permits a complete evaluation of tradeoffs between saturation drive current and subthreshold leakage current for a prescribed cycle time performance and operating temperature range. Constraints imposed by NTRS-projected package heat removal coefficients permit local clock rates to apply only within a macrocell whose size and total number are calculated using the stochastic distribution. Limits on the performance of CMOS logic circuits are shown to be imposed by total power dissipation which increases exponentially with clock frequency. Optimum supply

TABLE V
DECREASING V_{dd}/V_{to} CALCULATED FROM NTRS PROJECTIONS OF SUPPLY VOLTAGE AND SATURATION DRAIN CURRENT PER UNIT WIDTH

Year	'97	'99	'01	'03	'06	'09	'12
f_{clk} (GHz)	.75	1.2	1.4	1.6	2.0	2.5	3.0
NTRS V_{dd}	2.5	1.8	1.5	1.5	1.2	0.9	0.6
NTRS I_{sat} (NFET) ($\mu A/\mu m$)	600	600	600	600	600	600	600
NTRS I_{sat} (PFET) ($\mu A/\mu m$)	280	280	280	280	280	280	280
V_{dd}/V_{to} required (calculated using Transregional model)	2.9	3.0	3.0	2.6	2.55	2.38	2.43

TABLE VI
PARAMETER VALUES USED IN COMPARING THE TRANSREGIONAL MOSFET MODEL WITH HSPICE

Symbol	parameter	NFET Parameters	PFET parameters
V_{dd} (V)	Supply voltage	VDD=2.5	
L (μm)	Channel length	L=0.25 microns	
t_{ox} (A)	Gate oxide thickness	TOX=4.5E-9	
$\mu_{on,p}$ ($cm^2/V\text{-sec}$)	low field mobility	UO=305.46	UO=104.7
$v_{satn,p}$ (cm/sec)	Electron saturation velocity	VMAX=1E7	VMAX=8.3365E5
L_d (μm)	Lateral diffusion into channel from source/drain diffusion.	LD=0.025U	LD=0.025U
x_j (μm)	Junction depth	XJ=0.05U	XJ=0.05U
N_a (cm^{-3})	Substrate doping	NSUB=6.51E17	NSUB=6.51E17
ϕ_f (V)	Fermi potential in neutral bulk	PHI=0.456	PHI=0.456
N_g (cm^{-3})	Doping concentration in polysilicon gate	NGATE=1E19	NGATE=1E19
-	Type of gate material used: TPG=1, same as S/D diffusion	TPG=1	TPG=1
V_{to} (V)	Long-channel device threshold voltage	VTO=0.60	VTO=-0.60
θ (V^{-1})	Vertical field mobility degradation factor	THETA=0.1966	THETA=0.0679
ϵ_{jb} (F/cm ²)	Zero-bias bulk junction capacitance per unit area	CJ=1.9752E-3	CJ=1.9752E-3
-	Step function doping profile at junction boundaries	MJ=0.5	MJ=0.5
ζ_{jsw} (F/cm)	Zero-bias side-wall bulk junction capacitance per unit length	CJSW=7.9008E-11	CJSW=7.9008E-11
-	Bulk side-wall junction grading coefficient: 0.5 corresponds to a step junction.	MJSW=0.5	MJSW=0.5

voltages, device threshold voltages, and device channel widths corresponding to minimum total power are calculated out to the year 2012 for local and global critical paths. These projections are consistent with technology and cycle time forecasts by the NTRS. Limits on the cycle time performance imposed by power dissipation are projected for the same period. Concurrency-driven voltage scaling is projected to yield decreasing percentage reductions in total power.

APPENDIX A

See Table VI.

APPENDIX B DERIVATION OF (25)

This derivation makes two key assumptions (Fig. 16) in addition to the requirement of equal rise and fall times.

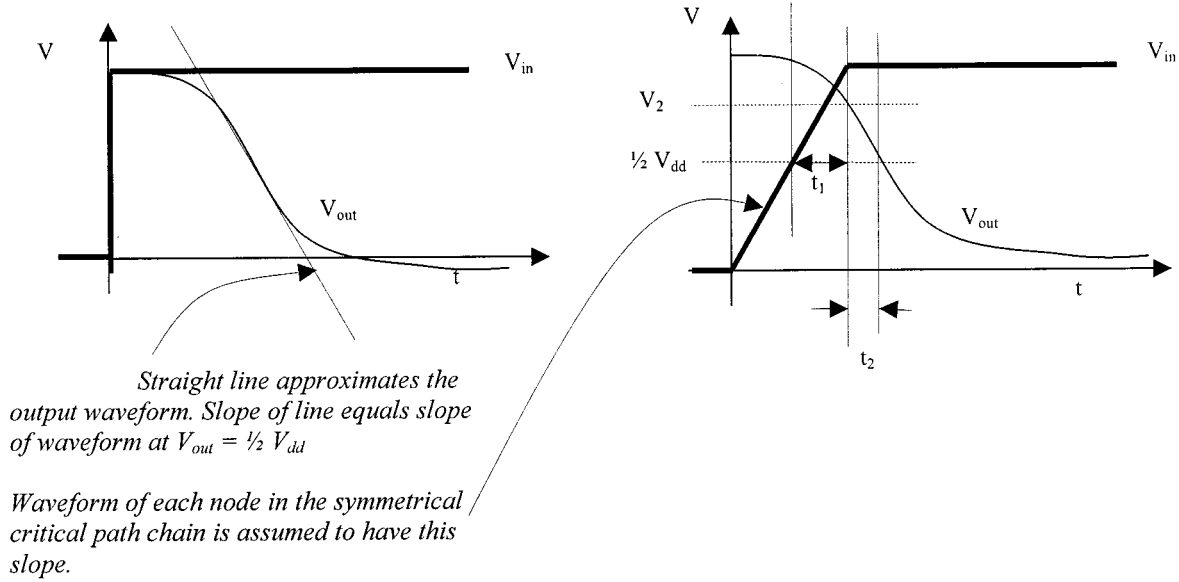


Fig. 16. Straight line approximates the output waveform. Slope of line equals slope of waveform at $V_{out} = \frac{1}{2} V_{dd}$. Waveform of each node in the symmetrical critical path chain is assumed to have this slope.

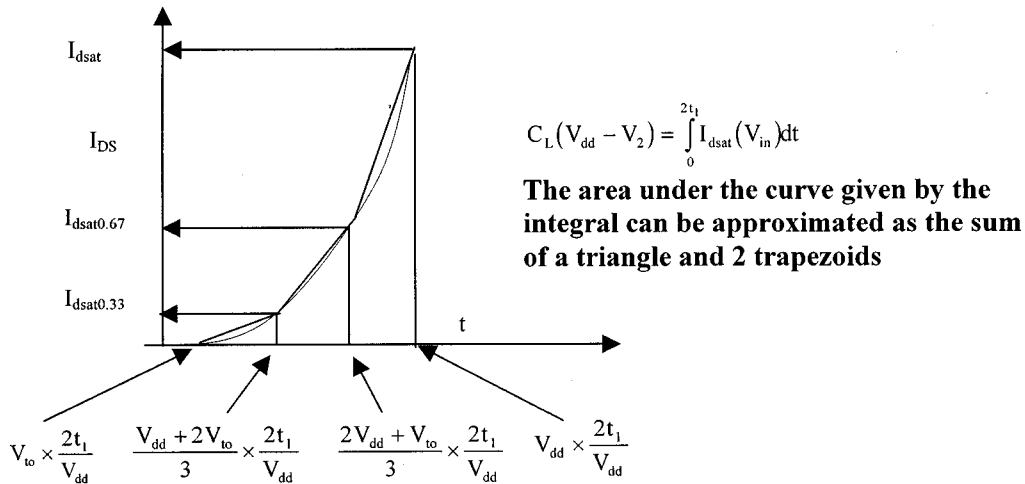


Fig. 17. $C_L(V_{dd} - V_2) = \int^{2t_1} I_{dsat}(V_{in}) dt$. The area under the curve given by the integral can be approximated as the sum of a triangle and two trapezoids.

- 1) The output waveform in response to a step input could be approximated as a straight line whose slope equals the slope of the waveform at half the transition
- 2) The waveform at each node in the chain of inverters has a slope equal to this straight line.

The time delay (24) between the input and output waveforms reaching $\frac{1}{2} V_{dd}$ is given as the sum of the following components:

$$t_{pd} = t_1 + t_2. \quad (B1)$$

The slope of the output waveform at $\frac{1}{2} V_{dd}$ in Fig. 16 is calculated using

$$C_L \frac{dV_{out}}{dt} = I_{dsat} \quad \text{or} \quad \left| \frac{dV_{out}}{dt} \right| = \frac{I_{dsat}}{C_L} \quad (B2)$$

and the waveform approximated as a straight line with the above slope has a base of width

$$t_1 = \frac{C_V V_{dd}}{2I_{dsat}}. \quad (B3)$$

The time delay t_2 , defined as the time taken for the output waveform to reach $\frac{1}{2} V_{dd}$ once the input waveform has completed its transition, requires calculation of V_2 , the output voltage at the time when the input completes its transition.

To calculate V_2

$$C_L \frac{dV_{out}}{dt} = I_{dsat}(V_{in}), \quad \text{as } V_{in} \text{ moves from zero to } V_{dd} \quad (B4)$$

integrating

$$C_L(V_{dd} - V_2) = \int_0^{2t_1} I_{dsat}(V_{in}) dt. \quad (B5)$$

The drain current of a MOSFET in the saturation region is between quadratic and linearly dependent on V_{gs} due to the presence of velocity saturation. Integrating the right-hand side (RHS) of (B5) would increase the complexity of the resulting expression for V_2 substantially. The integral in (B5) can be approximated as the sum of the areas of the triangle and the two trapezoids, whose bases are equal, as shown in Fig. 17: the area under the drain current in Fig. 17 is given by the sum of the areas of each of the solid figures that approximate the area under the curve

area of triangle:

$$\frac{1}{2} \frac{(V_{dd} - V_{t0})}{3} \frac{2t}{V_{dd}} I_{dsat.33} \quad (B6)$$

where $I_{dsat.33}$ is the drain saturation current when the gate input voltage is $V_{dd} + 2V_{t0}/3$

area of middle trapezoid:

$$\frac{1}{2} \frac{(V_{dd} - V_{t0})}{3} \frac{2t}{V_{dd}} (I_{dsat.33} + I_{dsat.66}) \quad (B7)$$

where $I_{dsat.66}$ is the drain saturation current when the gate input voltage is $2V_{dd} + V_{t0}/3$

area of right trapezoid:

$$\frac{1}{2} \frac{(V_{dd} - V_{t0})}{3} \frac{2t_1}{V_{dd}} (I_{dsat.66} + I_{dsat}). \quad (B8)$$

I_{dsat} is the drain saturation current when the gate input voltage is V_{dd}

Total area:

$$\frac{1}{2} \frac{(V_{dd} - V_{t0})}{3} \frac{2t_1}{V_{dd}} (2 \times I_{dsat.33} + 2 \times I_{dsat.66} + I_{dsat}). \quad (B9)$$

From (B5)

$$\begin{aligned} C_L(V_{dd} - V_2) \\ = \int_0^{2t_1} I_{dsat}(V_{in}) dt = \frac{1}{2} \frac{(V_{dd} - V_{t0})}{3} \frac{2t_1}{V_{dd}} \\ \cdot (2 \times I_{dsat.33} + 2 \times I_{dsat.66} + I_{dsat}) \end{aligned} \quad (B10)$$

using (B3), $2t_1 = (C_L V_{dd} / I_{dsat})$ in the above equation, and solving for V_2

$$V_2 = \frac{5V_{dd} + V_{t0}}{6} - \left(\frac{V_{dd} - V_{t0}}{3} \right) \left[\frac{I_{dsat.33}}{I_{dsat}} + \frac{I_{dsat.66}}{I_{dsat}} \right]. \quad (B11)$$

The time t_2 is calculated as the time required for the output to move from V_2 to $\frac{1}{2}V_{dd}$

$$C_L(V_2 - \frac{1}{2}V_{dd}) = t_2 I_{dsat} \quad (B12)$$

thus

$$\begin{aligned} t_2 = \frac{C_L}{I_{dsat}} \left[\frac{2V_{dd} + V_{t0}}{6} - \left(\frac{V_{dd} - V_{t0}}{3} \right) \right. \\ \left. \cdot \left(\frac{I_{dsat.33} + I_{dsat.66}}{I_{dsat}} \right) \right] \end{aligned} \quad (B13)$$

and

$$\begin{aligned} t_{pd} = t_1 + t_2 = \frac{C_L V_{dd}}{2I_{dsat}} \left[\frac{2V_{dd} + V_{t0}}{6} - \left(\frac{V_{dd} - V_{t0}}{3} \right) \right. \\ \left. \cdot \left(\frac{I_{dsat.33} + I_{dsat.66}}{I_{dsat}} \right) \right]. \end{aligned} \quad (B14)$$

REFERENCES

- [1] J. D. Meindl, "Low power microelectronics: Retrospect and prospect," *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.
- [2] A. Chandrakasan, S. Sheng, and R. Broderon, "Low power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473–484, Apr. 1992.
- [3] J. Burr *et al.*, "A 200 mV encoder-decoder circuit using stanford ultra low power CMOS," in *ISSCC Dig. Tech. Papers*, Feb. 1994, pp. 84–85.
- [4] M. Horowitz *et al.*, "Low power digital design," in *IEEE SLPE Dig. Tech. Papers*, Oct. 1994, vol. 1, pp. 8–11.
- [5] K. Itoh *et al.*, "Trends on lower power RAM circuit technologies," *Proc. IEEE*, vol. 83, pp. 524–539, Apr. 1995.
- [6] D. Liu *et al.*, "Trading speed for low power by choice of supply and threshold voltages," *IEEE J. Solid-State Circuits*, vol. 28, pp. 10–17, Jan. 1993.
- [7] The National Technology Roadmap for Semiconductors, *SIA Handbook*, 1997.
- [8] B. Agrawal, V. De, and J. Meindl, "Opportunities for scaling FET's for gigascale integration," in *Proc. 23rd ESSDERC*, Sept. 1993, pp. 919–926.
- [9] J. Davis, V. De, and J. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)—Parts I and III," *IEEE Trans. Electron Devices*, vol. 45, pp. 580–597, Mar. 1998.
- [10] C. Jacoboni *et al.*, "A review of some charge transport properties in silicon," *Solid-State Electron.*, vol. 20, pp. 77–85, 1977.
- [11] S. Garverick and C. Sodini, "A simple model for scaled MOS transistors that includes field-dependent mobility," *IEEE J. Solid-State Circuits*, vol. 22, pp. 111–114, Jan. 1987.
- [12] B. T. Murphy *et al.*, "Unified field effect transistor theory including velocity saturation," *IEEE J. Solid-State Circuits*, vol. 15, pp. 325–327, June 1990.
- [13] G. S. Halasz, "Performance trends in high-end processors," *Proc. IEEE*, vol. 83, pp. 20–36, Jan. 1995.
- [14] R. W. Keyes, "The wire limited logic chip," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 1232–1233, Dec. 1982.
- [15] B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.
- [16] J. Chen *et al.*, "Multilevel metal capacitance models for CAD design synthesis systems," *IEEE Electron Device Lett.*, vol. 13, pp. 32–33, Jan. 1992.
- [17] J. Uyemyra, *Circuit Design for CMOS VLSI*. Norwood, MA: Kluwer, 1994, p. 99.
- [18] R. K. Watts, *Submicron Integrated Circuits*. New York: Wiley, 1989, pp. 24–37.
- [19] J. Rabaey, *Digital Integrated Circuits—A Design Perspective*. Upper Saddle River, NJ: Prentice-Hall, 1996, pp. 133–134.
- [20] T. Sakurai and R. Newton, "Alpha power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, pp. 584–594, Apr. 1990.
- [21] —, "Delay models for series connected MOSFET structures," *IEEE J. Solid-State Circuits*, vol. 28, pp. 40–48, Jan. 1993.
- [22] C. Park *et al.*, "Reversal of temperature dependence of integrated circuits operating at very low voltages," in *IEEE IEDM 1995 Dig. Tech. Papers*, Dec. 1995, pp. 3.5.1–3.5.4.
- [23] S. Sze, *The Physics of Semiconductor Devices*, 2nd ed. New York: Wiley, 1981, p. 30, 451.
- [24] F. Najm, "Transition density: A new measure of activity in digital circuits," *IEEE Trans. Computer-Aided Design*, vol. 12, pp. 310–323, Feb. 1993.
- [25] P. E. Gronowski *et al.*, "High performance microprocessor design," *IEEE J. Solid-State Circuits*, vol. 33, pp. 676–686, May 1988.
- [26] J. D. Meindl, "The evolution of solid state circuits: 1958–1992–2000 (as reflected in the ISSCC digest of papers)," in *1993 ISSCC Commemorative Supplement*, Feb. 1993, pp. 23–26.

- [27] J. R. Pfiester, J. D. Shott, and J. D. Meindl, "Performance limits of CMOS ULSI," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 253–263, Feb. 1985.
- [28] D. J. Frank, P. Solomon, S. Reynolds, and J. Shin, "Supply and threshold voltage optimization for low power design," in *IEEE 1997 Int. Symp. Low Power Electronics and Design*, Aug. 1997, pp. 31–322.
- [29] A. P. Chandrakasan and R. W. Broderson, *Low Power Digital CMOS Design*. Norwood, MA: Kluwer, 1995.

Azeez J. Bhavnagarwala was born in Madras, India. He received the B.S. degrees (cum laude) in 1992 in electrical engineering from the Rensselaer Polytechnic Institute (RPI), Troy, NY, and from the Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, and the M.S. degree in electrical engineering from RPI in 1994. He is currently working toward the Ph.D. degree in electrical engineering at the Georgia Institute of Technology, Atlanta. His dissertation focuses on voltage scaling limits for CMOS logic and memory circuits.

He worked with integrated device technology (IDT) on a 1-Mb dual-port CMOS SRAM in 1996 and with LSI logic on interconnect and device modeling in 1999.

Blanca L. Austin received the B.S. degree in computer engineering from the University of Puerto Rico, R.U.M., Mayaguez, Puerto Rico, in 1986 and the M.S. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1987. She is currently working toward the Ph.D. degree in electrical engineering at the Georgia Institute of Technology.

She was a Member of Technical Staff at Bell Communications Research. Her research interests include deep-submicron low-power/high-performance MOS device design and characterization.

Keith A. Bowman received the B.S. degree in electrical engineering from North Carolina State University, Raleigh, in 1994 and the M.S. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1995. He is currently working toward the Ph.D. degree in electrical engineering at the Georgia Institute of Technology.

His research interests include high-performance/low-power device and circuit optimizations for future generations of technology.

James D. Meindl (LF'98) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Carnegie-Mellon University, Pittsburgh, PA, in 1955, 1956, and 1958, respectively.

He is presently the Director of the Joseph M. Pettit Microelectronics Research Center and the Pettit Chair Professor of Microelectronics at the Georgia Institute of Technology, Atlanta, GA. Previously, he served from 1986 to 1993 as Senior Vice President for Academic Affairs and Provost at the Rensselaer Polytechnic Institute, Troy, NY. From 1967 to 1986, he was with Stanford University, Stanford, CA, where he was the John M. Fluke Professor of Electrical Engineering, Associate Dean for Research in the School of Engineering, Director of the Center for Integrated Systems, Director of the Electronics Laboratories, and Founding Director of the Integrated Circuits Laboratory. He is a cofounder of Telesensory Systems, Inc. and the principal manufacturer of electronic reading aids for the blind. He served as a Member of the Board from 1971 through 1984. From 1965 to 1967, he was Founding Director of the Integrated Electronics Division, U.S. Army Electronics Laboratories, Fort Monmouth, NJ. He is the author of *Micropower Circuits* and over 500 technical papers on ultralarge scale integration, integrated electronics, and medical electronics and Editor of *Brief Lessons in High Technology*, which elucidates the most important economic event of our lives, the emergence of the information society. His major contributions have been: new medical instruments enabled by custom-integrated electronics; projections and codification of the hierarchy of physical limits on integrated electronics; and leadership in creation of academic environments promoting high-quality teaching and research.

Dr. Meindl is a Life Fellow of the American Association for the Advancement of Science. He is a member of the American Academy of Arts and Sciences and the National Academy of Engineering and its Academic Advisory Board. He received the IEEE Third Millennium Medal, the 1999 SIA University Research Award, the 1997 Hamerschlag Distinguished Alumnus Award from Carnegie-Mellon University, and the 1991 Benjamin Garver Lamme Medal from ASEE. He was the recipient of the 1990 IEEE Education Medal "for establishment of a pioneering academic program for the fabrication and application of integrated circuits" and the recipient of the 1989 IEEE Solid-State Circuits Medal for contributions to solid-state circuits and solid-state circuit technology. At the 1988 IEEE International Solid-State Circuits Conference, he received the Beatrice K. Winner Award. In 1980, he was the recipient of the IEEE Electron Devices Society's J. J. Ebers Award for his contributions to the field of medical electronics and for his research and teaching in solid-state electronics. From 1970 to 1978, he and his students received five outstanding paper awards at the IEEE International Solid-State Circuits Conferences, along with one received at the 1985 IEEE VLSI Multilevel Interconnections Conference.