CMOS System-on-a-Chip Voltage Scaling beyond 50nm

Azeez J Bhavnagarwala, Blanca Austin, Ashok Kapoor[‡] and James D Meindl
Microelectronics Rserch. Cntr. and School of Elec. and Comp. Engr., Georgia Institute of Technology, Atlanta GA 30332
[‡]*LSI Logic Corporation, Milpitas CA 95035*

Abstract[†]

The limits on CMOS energy dissipation imposed by subthreshold leakage currents and by wiring capacitance are investigated for CMOS generations beyond 50nm at NTRS projected local and global clock rates for high performance processors. Physical short-channel MOSFET models that consider high-field effects, threshold voltage roll-off and reverse subthreshold swing roll-off are employed in tandem with stochastic interconnect distributions to calculate optimal supply voltage, threshold voltage and gate sizes that minimize total CMOS power dissipation by exploiting trade-offs between saturation drive current and subthreshold leakage current and between device size and wiring capacitance. CMOS power dissipation at its lower limit, increases exponentially with clock frequency imposing limits on performance set by heat removal. Heat removal constraints at high local clock rates, limiting the average wire length and device size within a local zone of synchrony, or macrocell, in a short-wire cellular array architecture are used to project the maximum macrocell size and count for generations beyond 100nm.

1. Introduction

The supply voltage for future gigascale integrated systems are projected to scale to 0.37V for the 35nm, 17GHz generation [1] to reduce electric field strengths and also power dissipation (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 bulk Si CMOS logic circuits will be to meet the projected performances given the competing requirements of high performance and low standby power at low voltages [1,2,3] in the presence of threshold voltage reductions due to short-channel effects and subthreshold swing increases due to the 2D electrostatic charge coupling between gate and source/drain terminals of the MOSFET. A methodology [4] simultaneously considering the device, circuit and system levels of the design hierarchy and distinguishing local and global clock rates, is employed to minimize total power dissipated from a static CMOS critical path gate during a clock cycle. This methodology assumes a realistic environment of chip size, logic gate count, clock frequency, wiring capacitance, critical path depth and range of operating temperature. This analysis uses physical and stochastic models, verified by HSPICE, MEDICI and actual microprocessor implementations to investigate opportunities to scale V_{dd} to the optimal point corresponding to the limits of CMOS power dissipation where leakage power balances switching power dissipation, and when device capacitance balances wiring capacitance. The analysis considers Retrograde Doped (RD) (Fig.2) MOSFETs – the bulk Si alternative to a Uniformly Doped (UD) MOSFET that promises, higher performance and superior scalability [5] (Fig 3).

2. Circuit and Device Models

The performance of a generic CMOS processor is modeled assuming a global critical path of 15 [6], 2-way NAND stages, each stage driving average wire lengths (Fig 4). Average wire lengths, in units of gate pitches, are determined (Table 1) from stochastic interconnect distributions [7], derived recursively using Rent's rule, and verified for an actual microprocessor in Fig. 5. In logic-intensive CMOS chips, packing densities are interconnect limited [8] where the effective size of a gate is determined by its wireability [9]. The gate pitch is estimated from NTRS projections for microprocessor chip size, and logic transistor count after discounting the extrapolated increases in cache size and cache area for high performance processors (Fig 6). Assuming equal interconnect cross-sectional dimensions, and that neighboring wiring planes in a multi-level network provide an approximate ground plane, total capacitance per unit length, including fringing effects, is estimated using analytical models in [10]. Device performance is modeled using compact low-voltage Transregional MOSFET models [11,12] (Figs 7,8,9) that predict circuit performance in the sub-threshold, saturation and linear regions of operation providing continuous and smooth transitions across region boundaries. High field-effects on carrier mobility are incorporated by adopting the mobility reduction model in [13]. Smoothness and continuity of the drain current expressions in the triode, saturation and the subthreshold regions are obtained by requiring differentiability and continuity of the product of the effective mobility and the areal charge density of inversion layer carriers. Low field mobility dependence on temperature and doping concentration is estimated using empirical models reported in [14]. The doping profile for the RD structure is selected as one that yields the smallest depletion depth, corresponding to the least DIBL effects for a given V_{to} and gate oxide thickness [15]. Increases in leakage current due to DIBL (Drain Induced Barrier Lowering) effects are calculated using 2D subthreshold models [6] that accurately predict the threshold voltage roll-off and subthreshold swing increase (Fig 10) dependence on supply voltage, device geometries and doping profile. The 2-way NAND gate, as a basic circuit building block in the critical path, has a performance that parallels that of any other circuit actually used in processor critical paths in reflecting technology improvements [16]. The improved delay dependence on fan-in at short channel lengths [17] due to a smaller reduction in the saturation drain current with a rise in the source voltage of the topmost series-connected MOSFET is modeled physically

[†] This work was supported by the Defense Advanced Research Project Agency (Contract: F3361595C1623) and the Semiconductor Research Corporation (SJ-374-002)

by calculating the fractional reduction of the normalized saturation drain current for the series-connected struc. [18].

3. Minimum Power CMOS Random Logic Networks

Power drain of a static CMOS gate is minimized by scaling the supply voltage while meeting the performance required by scaling the threshold voltage and increasing the channel widths until further decrease in threshold voltage, increases total power due to a dominating static component [3] (Figs 10, 11) and further increases in device size increase total power due to larger gate sizes [19] (Fig 12). Optimal supply voltage (Fig 13), device threshold voltage and gate sizes are calculated corresponding to a simultaneous solution at these minima (Table 2). For a given wiring load, the performance of a static CMOS gate increases asymptotically with increasing (W/L) ratios, with gate delays reaching past the knee of the asymptotic dependence of delay on channel width. for wiring capacitance less than or equal to 40% of the total load capacitance. This point corresponds to minimum power with respect to gate size where further increases in gate size increases power linearly while permitting only asymptotic reductions in supply voltage. Critical path gates clocked at high local frequencies are assumed to be only 5 stages long and drive wire lengths averaged within a macrocell of a 'short-wire' cellular array architecture (Fig 14). Assuming gates are sized so that wiring capacitance is 40% of the total load, the cell count (Table 3) is calculated using the stochastic interconnect distribution by imposing a maximum heat removal coefficient of 50 W/cm^2 on the average wire length of the cell, calculated using the stochastic distribution. Total CMOS power increases exponentially (Fig 15) for a given generation, with increases in clock frequency due to an exponential rise in the supply voltage necessary to meet shrinking cycle times and the accompanying increases in leakage current due to threshold voltage reductions and subthreshold swing increases. The maximum heat removal coefficients of the package thus impose limits on CMOS performance.

4. Summary and Conclusions

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 roll-off and subthreshold swing roll-up models in tandem with stochastic wiring distributions. Optimum supply voltages, device threshold voltages, and device channel widths corresponding to minimum total power are calculated out to year 2014 for local and global critical paths. These projections are consistent with technology and cycle time forecasts by the NTRS. Limits on the performance of CMOS logic circuits are shown to be imposed by total power dissipation which increases exponentially with clock frequency. Limits on the cycle time performance imposed by power dissipation are projected for the same period. 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 stocahstic distribution.

5. References

1] The 1997 NTRS, Semiconductor Industry Association, Dec 1997

2] J D Meindl, *'Low Power Microelectronics - Retrospect and Prospect'*, Proceedings of the IEEE, Vol. 83, No 4 Apr 1995, pg 619.

3] J Burr and J Shott, *'A 200mV Encoder-Decoder circuit Using Stanford Ultra Low Power CMOS'* ISSCC Dig Tech Papers, Feb 1994, pp 84-85.

4] A Bhavnagarwala, V. De, B Austin and J Meindl, *"Circuit Techniques for Low Power CMOS GSI"*, IEEE ISLPED, Aug 1996 Dig, pp 193-197

5] B Agrawal, V. De and J Meindl, *"Opportunities for Scaling FET's for Gigascale Integration"*, Proceedings of the 23rd ESSDERC, Sept 1993, pp 919 – 926.

6] P E Gronowski et al, "High performance microprocessor design*", IEEE Journal of Solid State Circuits*, Vol 33, No 5, pp 676-686, May 1988.

7] J Davis, V. De & J. Meindl, *"A stochastic wire-length distribution for gigascale integration (GSI) – parts I & II"*, IEEE Transactions on Electron Devices, Vol 45, No. 3, pp580-597, March 1998

8] R W Keyes, *"The Wire Limited Logic Chip"*, IEEE JSSC, Vol SC-17, Dec 1982, pp 1232-1233

9] B Bakoglu, *"Circuit Interconnections and Packaging for VLSI"*, Addison Wesley, 1990

10] J Chern et al, *"Multilevel Metal Capacitance Models for CAD Design Synthesis Systems"* IEEE EDL Vol 13, No 1, Jan 1992, pg 32.

11] R Swanson & J Meindl, *"Ion-Implanted Complentary MOS Transistors in Low Voltage Circuits*", IEEE JSSC, Vol. SC-7, pp. 146-153, Apr. 1972

12] B. Austin, K. Bowman, Xinghai Tang, and J. D. Meindl, "*A Low Power Transregional MOSFET Model for Complete Power-Delay Analysis of CMOS Gigascale Integration (GSI),"* Proc. of the 11th Annual IEEE Intl. ASIC Conf., pp. 125-129, Sept. 1998

13] C Sodini, P Ko and J Moll, *'The Effect of High Fields on MOS Device and Circuit Performance'*, IEEE TED, Vol ED-31, No 10, October 1984, pp 1386

14] C Jacoboni et al, '*A review of some charge transport properties in silicon'*, Solid State Electronics, No 20, Vol 77, 1977

15] B Agrawal V. De and J Meindl, *"Device Parameter Optimization for Reduced Short Channel Effects in Retrograde Doped MOSFETs"*, IEEE TED, Vol 43, No 2, Feb 1996, pg 365

16] G Sai Halasz, '*Performance Trends in High-end Processors,*' Proceedings of the IEEE, Vol 83, Jan 1995, pp 20-36

17] T Sakurai & R Newton, *"Delay Models for Series Connected MOSFET Structures"* IEEE JSSC, Vol 28, No 1, Jan 1993, pg 40

18] A Bhavnagarwala, B Austin, J Meindl, *"Minimum Supply Voltage for bulk Si CMOS GSI"*, IEEE ISLPED, Aug 1998 Dig, pp 100-103

19] A Chandrakasan, S Sheng and R Broderson, *'Low-Power CMOS Digital Design*', IEEE JSSC Vol 27, No 4, April 1992, pp 473-484

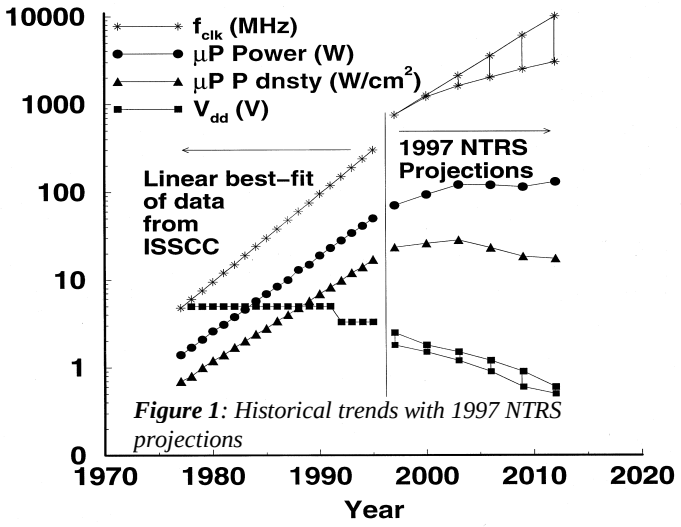


***Figure 1**: Historical trends with 1997 NTRS projections*

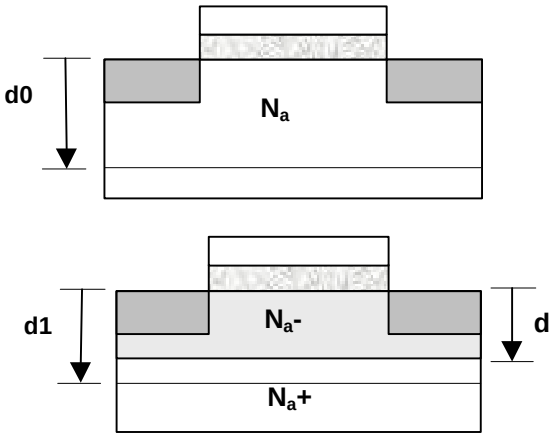


***Figure 2:** Shallow junction Uniform Doped (UD) and Retrograde Doped (RD) MOSFETs.*

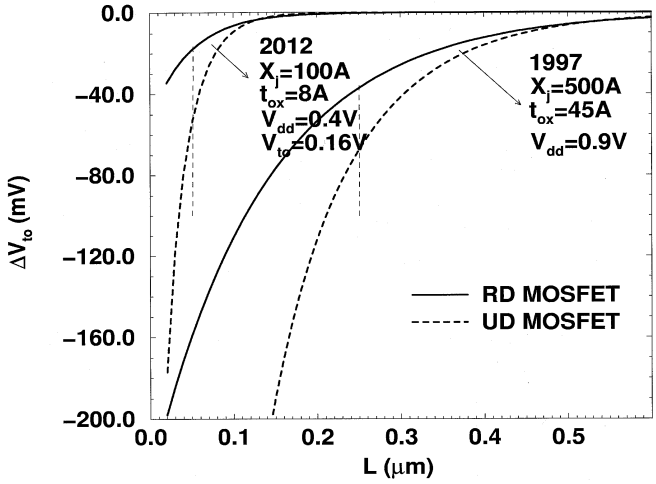


***Figure 3:** Calculated V_{to} roll-off for bulk Si at NTRS projected gate oxide thickness [6]*

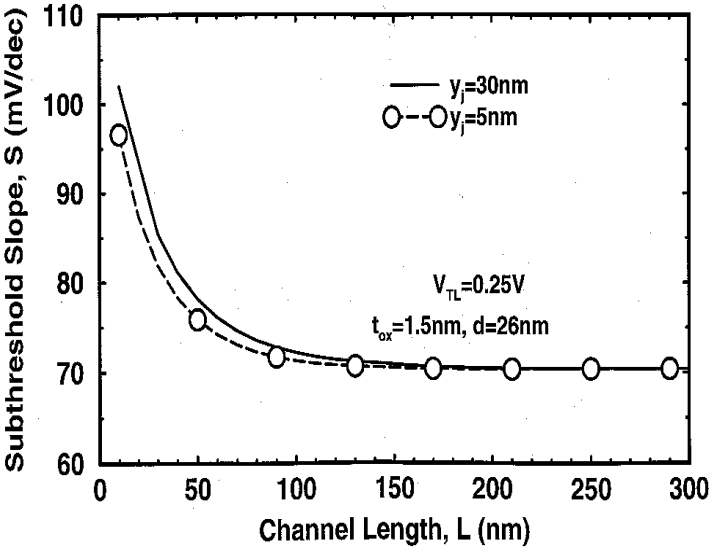
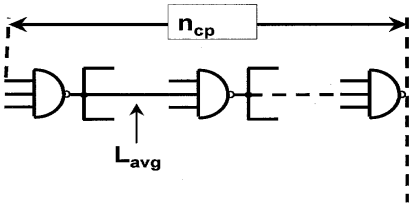


***Figure 4 :** Subthreshold swing increases accompany threshold voltage reductions increasing stand-by currents substantially*



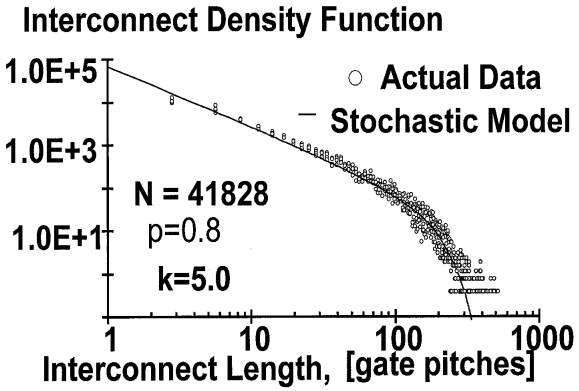


***Figure 5 [7]:** Stochastic wiring distribution comparison with an actual microprocessor implementation. The distribution is used to calculate the average interconnect length between two logic gates*

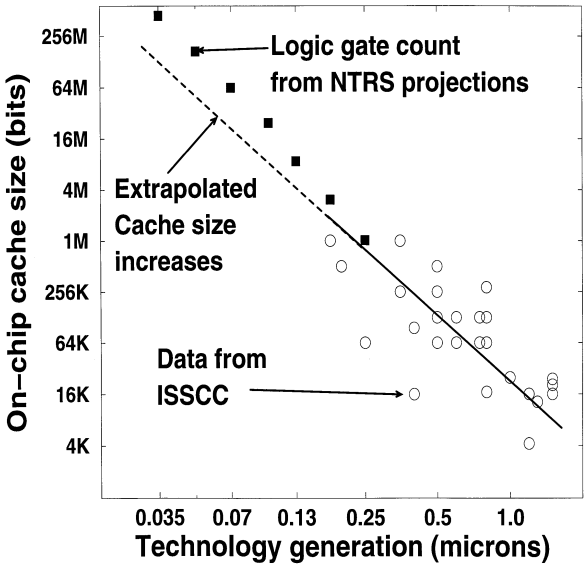


Figure 6: *Cache size extrapolations to discount SRAM cell transistors from total transistor count when calculating average wire length of a logic network*

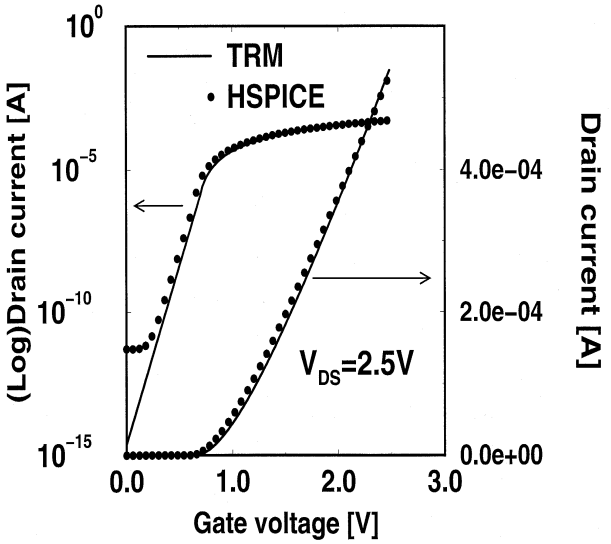


Figure 7 *: Comparison of 0.25 micron CMOS HSPICE gate characteristics with the Transregional model (TRM). W=0.5mm*

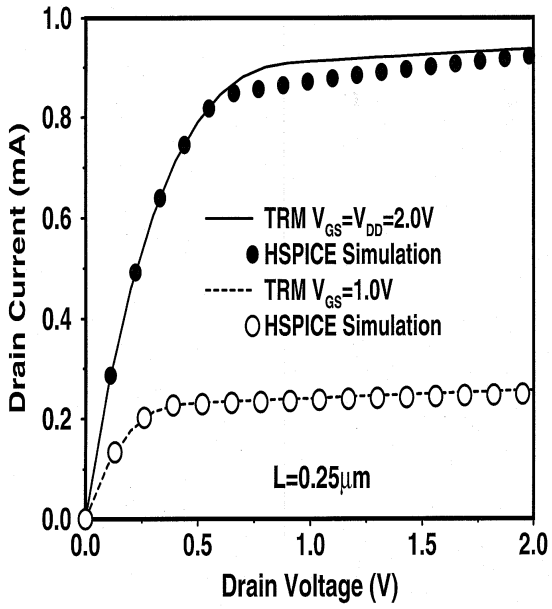


Figure 8 : *Comparison of 0.25 micron CMOS HSPICE drain characteristics with the Transregional model (TRM). W=0.5mm*

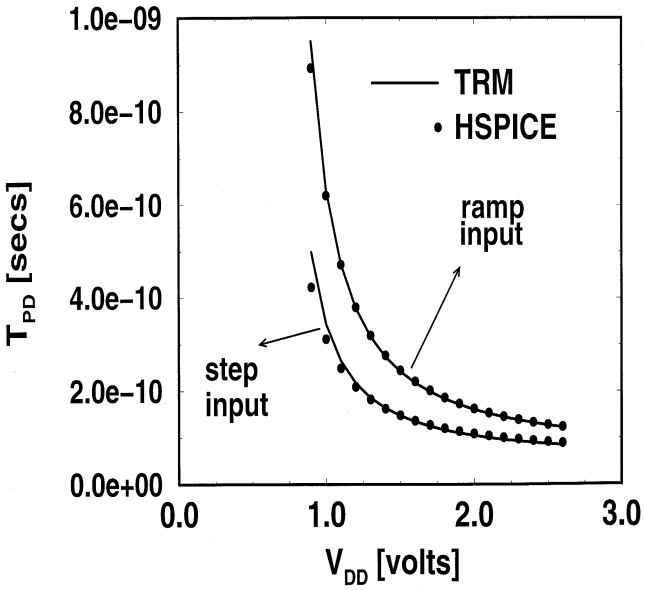


Figure 9 : *Comparison of 0.25 micron CMOS HSPICE simulations with propagation delay models used from [4]*

Year	'97	'99	'02	'05	'08	'11	'14
F(mm)	.25	.18	.13	.10	.07	.05	.035
Tox (A)	45	32	22	15	11	8	6
f_{clk}(GHz)	.75	1.2	1.6	2.0	2.5	3.0	3.7
V_{topt} (V)	0.22	0.21	0.2	0.18	0.17	0.16	0.16
V_{ddopt} (V)	1.23	1.01	0.91	0.72	0.64	0.52	0.41
- DV_{TO} (V)	103	95	88	75	54	39	42
DS (mV/dec)	2.0	2.2	2.7	3.2	3.2	3.0	5.0
P_{total}(mW)	17.1	14.3	11.4	8.6	6.3	4.9	3.8

Table-2: Optimal $V_{dd,}$ V_{topt}, $W/L_{n,p}$ for across-chip global clock rates. NTRS projected gate oxide thickness are assumed.

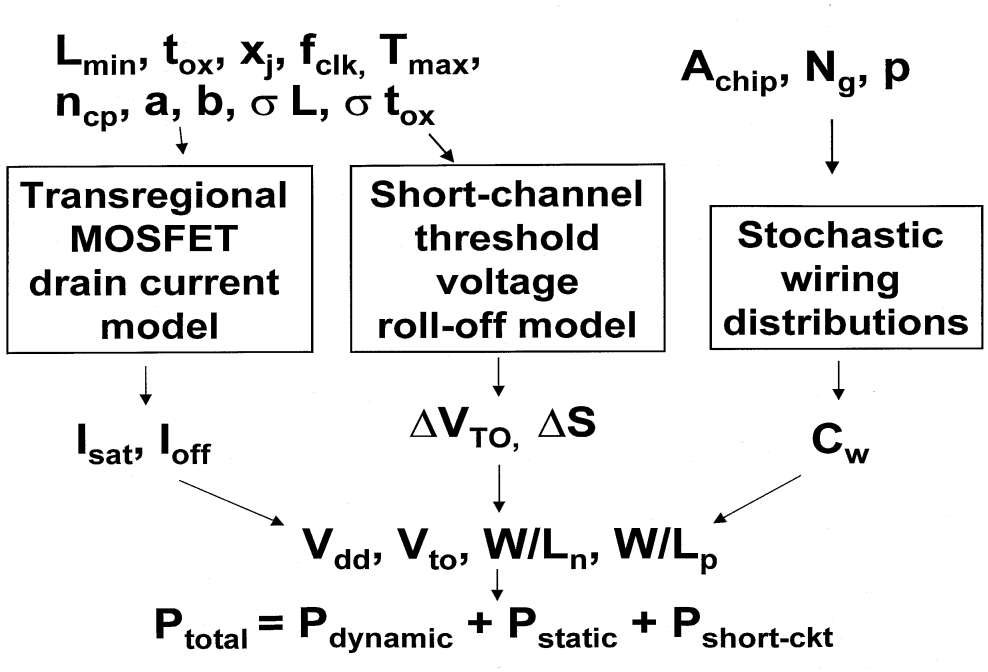
Year	F (mm)	Chip size, cm^2	N_{gates} $\times 10^6$	Cw (fF)
1997	*0.25*	*3.0*	*1.07*	33.4
1999	*0.18*	*3.4*	*3.1*	24.3
2002	*0.13*	*4.3*	*9.1*	18.1
2005	*0.10*	*5.2*	*25.7*	15.5
2008	*0.07*	*6.2*	*66.7*	12.1
2011	*0.05*	*7.5*	*177.5*	9.4
2014	*0.035*	*9.0*	*465.5*	7.9

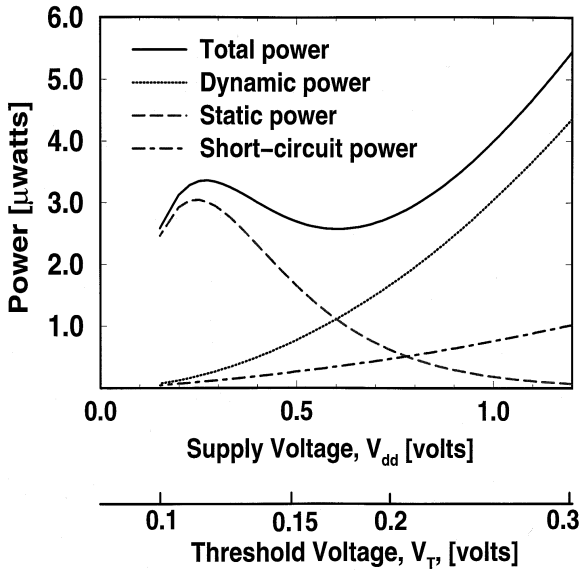
Table 1: *Average wiring capacitance estimates for NTRS generations using the stochastic interconnect distribution.*

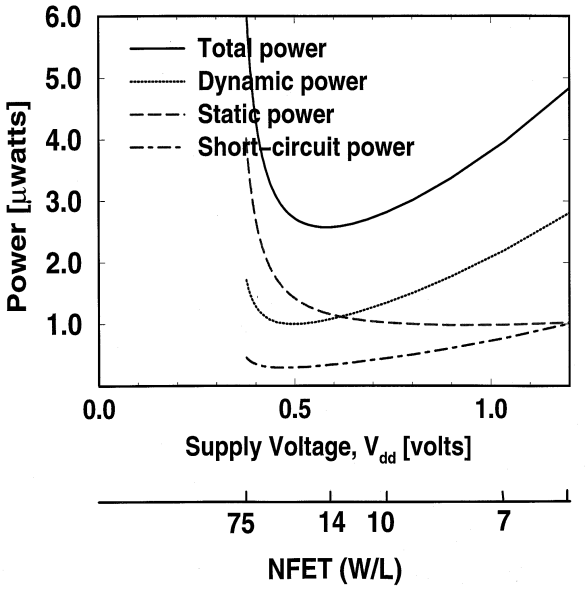
Yr	'05	'08	'11	'14
F(μm)	.10	.07	.05	.035
T_{ox} (A)	15	11	8	6
F_{clk} (GHz)	3.5	6.0	10.0	16.9
C_w (fF)	15.6	11.1	7.9	5.7
N_{cells}	72	266	1105	4412
V_{ddopt}	1.05	0.75	0.55	0.51
V_{topt}	0.19	0.18	0.16	0.14

Table-3: *Average wire lengths and wiring capacitance imposed by heat removal for the sub-100nm generations. Size and number of macrocells are calculated using the stochastic wiring distribution [7]* Q=50W/cm^2

Figure 10: *Physical drain current and short channel MOSFET threshold voltage roll-off models are used with stochastic interconnect distributions, to project optimal critical path gate designs minimizing total power dissipated by CMOS logic circuits for each NTRS technology generation.*







L (min feature size)= 50nm
f_{clk} (local clock rate) = 10GHz
t_{ox} (gate oxide thickness) = 8 Å
a (% switching activity)= 0.05
b (clock skew) = 0.9
n_{cp} (logic depth) = 5
C_w (average wire cap)= 4.4fF
f_{in} (average fan-in) = 2
f_{out} (average fan-out) = 2
T_{MAX} (maximum temperature) = 400°K
P (Rent's exponent) = 0.6
V_{ddopt} (optimal Vdd) = 0.6V
V_{topt} (optimal Vto) = 0.17V
$(W/L)_n$ (optimal NFET W/L) = 14
$(W/L)_p$ (optimal PFET W/L) = 16

***Figure 11 &12** : Total power dissipation and its component's dependence on supply voltage, threshold voltage and NFET channel width. PFET channel width is calculated for equal rise and fall times.*

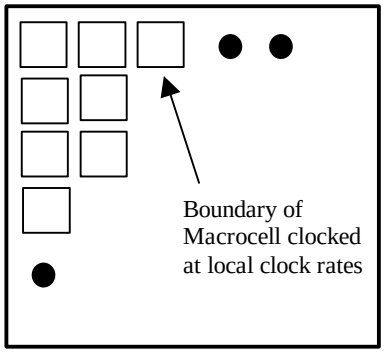
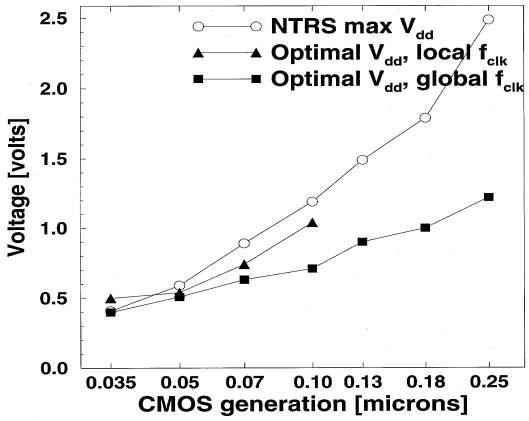


***Figure 13(at left):** A short-wire cellular array architecture with local and global clock frequencies where local clocks apply only within the boundary of a macrocell*



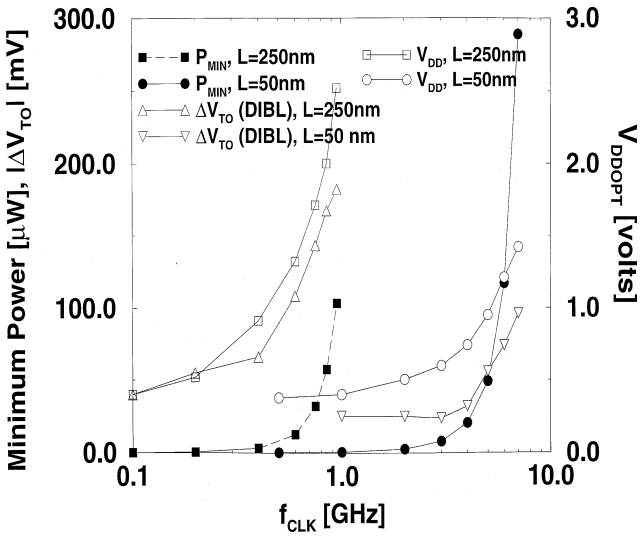


***Figure 14 (at left):** Optimal V_{dd} and NTRS projections*

***Figure 15 (above):** Exponential increase in power with clock frequency impose limits on CMOS performance*