

# Low Power Microelectronics: Retrospect and Prospect

JAMES D. MEINDL, FELLOW, IEEE

*The era of low power microelectronics began with the invention of the transistor in the late 1940's and came of age with the invention of the integrated circuit in the late 1950's. Historically, the most demanding applications of low power microelectronics have been battery operated products such as wrist watches, hearing aids, implantable cardiac pacemakers, pocket calculators, pagers, cellular telephones and prospectively the hand-held multi-media terminal. However, in the early 1990's low power microelectronics rapidly evolved from a substantial tributary to the mainstream of microelectronics. The principal reasons for this transformation were the increasing packing density of transistors and increasing clock frequencies of CMOS microchips pushing heat removal and power distribution to the forefront of the problems confronting the advance of microelectronics.*

*The distinctive thesis of this discussion is that future opportunities for low power gigascale integration (GSI) will be governed by a hierarchy of theoretical and practical limits whose levels can be codified as: 1) fundamental, 2) material, 3) device, 4) circuit, and 5) system. The three most important fundamental limits on low power GSI are derived from the basic physical principles of thermodynamics, quantum mechanics, and electromagnetics. The key semiconductor material limits are determined by carrier mobility, carrier saturation velocity, breakdown field strength and thermal conductivity, and the prime material limit of an interconnect is imposed by the relative dielectric constant of its insulator. The most important device limit is the minimum channel length of a MOSFET, which in turn determines its minimum switching energy and intrinsic switching time. Channel lengths below 60 nm for bulk MOSFET's and below 30 nm for dual gate SOI MOSFET's are projected. Response time of a canonical distributed resistance-capacitance network is the principal device limit on interconnect performance. To insure logic level restoration in static CMOS circuits, a minimum allowable supply voltage is about  $4kT/q$ . For a conservative 0.1  $\mu\text{m}$  CMOS technology and 1.0 V supply voltage, the minimum switching energy of a ring oscillator stage is about 0.1 fJ and the corresponding delay time is less than 5.0 ps. Five generic system limits are set by: 1) the architecture of a chip, 2) the power-delay product of the CMOS and interconnect technology used to implement the chip, 3) the heat removal or cooling capacity of the packaging technology, 4) the cycle time requirements imposed on the chip and 5) its physical size.*

*To date, all microchips have been designed to dissipate the entire amount of electrical energy transferred during a binary switching transition. However, new approaches based on the second law*

*of thermodynamics point the way to recycle switching energy by avoiding the erasure of information and switching under quasi-equilibrium conditions. Adiabatic computing technology offers promise of significant new advances in low power microelectronics.*

*Practical limits are elegantly summarized by Moore's Law which defines the exponential rate of increase with time of the number of transistors per chip. One billion transistor chips are projected for the year 2000 and 100 billion transistor chips are projected before 2020 by joining the results of the analyses of theoretical and practical limits through definition of the chip performance index as the quotient of the number of transistors per chip and the power delay product of the corresponding technology.*

## I. INTRODUCTION

Low power microelectronics was conceived through the invention of the transistor in 1947 and enabled by the invention of the integrated circuit in 1958. Throughout the following 37 years, microelectronics has advanced in productivity and performance at a pace unmatched in technological history. Minimum feature size  $F$  has declined by about a factor of 1/50; die area  $D^2$  has increased by approximately 170 times; packing efficiency  $PE$ , defined as the number of transistors per minimum feature area has multiplied by more than a factor of 100 so that the composite number of transistors per chip  $N = F^{-2} \cdot D^2 \cdot PE$  has skyrocketed by a factor of about  $50 \times 10^6$ , while the price range of a chip has remained virtually unchanged and its reliability has increased [1]. An inextricable concomitant advance of low power microelectronics has been a reduction in the switching energy dissipation  $E$  or power-delay product  $Pt_d = E$  of a binary transition by approximately  $1/10^5$  times. Consequently, as the principal driver of the modern information revolution, the ubiquitous microchip has had a profound and pervasive impact on our daily lives. Therefore, it is imperative that we gain as deep an understanding as possible of where we have been and especially of where we may be headed with the world's most important technology.

Almost two decades ago Gordon Moore of Intel Corporation observed that the number of transistors per chip had been doubling annually for a period of 15 years [2]. This astute observation has become known as "Moore's Law." With a reduction of the rate of increase to about 1.5 times per year, or a quadrupling every three years, Moore's Law has remained through 1994 an accurate description

Manuscript received November 14, 1994; revised January 12, 1995. This work was supported in part by SRC Contract 93-SJ-374.

The author is with the School of Electrical and Computer Engineering and Microelectronics Research Center, Georgia Institute of Technology, Atlanta, GA 30332-0269 USA.

IEEE Log Number 9409346.

0018-9219/95\$04.00 © 1995 IEEE

Hierarchical Matrix of Limits

	Theoretical	Practical
5. System		
4. Circuit		
3. Device		
2. Material		
1. Fundamental		

Fig. 1. Hierarchical matrix of limits on GSI.

of the course of microelectronics. This discussion defines a corollary of Moore's Law which asserts that "future opportunities to achieve multi-billion transistor chips or gigascale integration (GSI) in the 21st century will be governed by a hierarchy of limits." The levels of this hierarchy can be codified as 1) fundamental, 2) material, 3) device, 4) circuit, and 5) system [3]. At each level there are two different kinds of limits to consider, theoretical and practical. Theoretical limits are informed by the laws of physics and by technological invention. Practical limits, of course, must comply with these constraints but must also take account of manufacturing costs and markets. Consequently, the path to GSI will be governed by a hierarchical matrix of limits as illustrated in Fig. 1, which emphasizes the structure of the hierarchy.

Following this introduction, Section II provides a brief retrospective view of low power microelectronics in which the antecedents of many current innovations are cited. Then, Section III treats the most important theoretical limits associated with each level of the hierarchy introduced in the preceding paragraph. In order to elucidate opportunities for low power microelectronics, many of these limits are represented by graphing the average power transfer  $P$  during a binary switching transition versus the transition time  $t_d$ . For logarithmic scales, diagonal lines in the  $P$  versus  $t_d$  plane represent loci of constant switching energy. Limits imposed by interconnections are represented by graphing the square of the reciprocal length of an interconnect  $(1/L)^2$  versus the response time  $\tau$  of the corresponding circuit. For logarithmic scales, diagonal lines in the  $(1/L)^2$  versus  $\tau$  plane represent loci of constant distributed resistance-capacitance product for an interconnect. The twin goals of low power microelectronics are to drive both the  $P$  versus  $t_d$  and the  $(1/L)^2$  versus  $\tau$  loci toward the lower left corners of their allowable zones of operation reflecting switching functions consuming minimal power and time, and communication functions covering maximal distance in minimal time.

Virtually all previous and contemporary microchips dissipate the entire amount of electrical energy transferred during a binary switching transition. This assumption is made in deriving the hierarchy of limits represented in the  $P$  versus  $t_d$  plane in Section III. However, in Section IV this stipulation is removed in a brief discussion of a new hierarchy of limits on quasi-adiabatic switching operations that recycle, rather than dissipate, a fraction of the energy transferred during a binary switching transition [4], [5].

In Section V practical limits are compactly summarized in a sequence of plots of minimum feature size, die edge, packing efficiency, and number of transistors per chip versus calendar year. Then the results of the discussions of theoretical and practical limits are joined by defining the most important single metric that indicates the promise of a technology for low power microelectronics, and that is the chip performance index or CPI which equals the quotient of the number of transistors per chip and the associated switching energy or  $CPI = N/Pt_d$ . Section VI concludes with a speculative comment on a paramount economic issue.

## II. BACKGROUND

The genesis of low power microelectronics can be traced to the invention of the transistor in 1947. The elimination of the crushing needs for several watts of heater power and several hundred volts of anode voltage in vacuum tubes in exchange for transistor operation in the tens of milli-watts range was a breakthrough of virtually unparalleled importance in electronics. The capability to fully utilize the low power assets of the transistor was provided by the invention of the integrated circuit in 1958. Historically, the motivation for low power electronics has stemmed from three reasonably distinct classes of need [6]–[12]: 1) the earliest and most demanding of these is for portable battery operated equipment that is sufficiently small in size and weight and long in operating life to satisfy the user; 2) the most recent need is for ever increasing packing density in order to further enhance the speed of high performance systems which imposes severe restrictions on power dissipation density; and 3) the broadest need is for conservation of power in desktop and deskside systems where a competitive life cycle cost-to-performance ratio demands low power operation to reduce power supply and cooling costs. Viewed in toto, these three classes of need appear to encompass a substantial majority of current applications of electronic equipment. Low power electronics has become the mainstream of the effort to achieve GSI.

The earliest and still the most urgent demands for low power electronics originate from the stringent requirements for small size and weight, long operating life, utility, and reliability of battery operated equipment such as wrist watches, pocket calculators and cellular phones, hearing aids, implantable cardiac pacemakers, and a myriad of portable military equipments used by individual foot soldiers [6], [7]. Perhaps no segment of the electronics industry has a growth potential as explosive as that of the personal digital assistant (PDA) which has been characterized as a combined pocket cellular phone, pager, e-mail terminal, fax, computer, calendar, address directory, notebook, etc. [8]–[12]. To satisfy the needs of the PDA for low power electronics, comprehensive approaches are proposed that include use of the lowest possible supply voltage coupled with architectural, logic style, circuit, and CMOS technology optimizations [8]–[12]. The antecedents of these concepts are strikingly evident in publications from the 1960's [6], [7], in which several critical principles of low power design

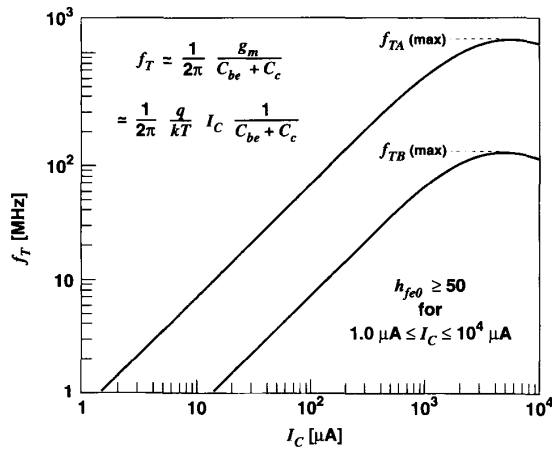


Fig. 2. Transistor gain-bandwidth product versus quiescent collector current.  $V_{CE} = 3.0$  V for transistors A and B [7].

were formulated and codified [7]. The first of these was simply to use the lowest possible supply voltage, preferably a single cell battery. The second guideline was to use analog techniques wherever possible particularly in order to avoid the large standby power drain of then available bipolar digital circuits, although the micropower potential of CMOS was clearly articulated by G. Moore *et al.* in 1963 [13].

A third key principle of micropower design that was convincingly demonstrated quite early is the advantage of selecting the smallest geometry, highest frequency transistors available to implement a required circuit function, e.g., a wideband amplifier, and then scaling down the quiescent current until the transistor gain-bandwidth product  $f_T$  just satisfies the relevant system performance requirements. (The manifestation of this concept in current CMOS technology is to seek the available technology with the smallest minimum feature size in order to reduce the capacitance that must be charged/discharged in a switching transition.) Bipolar transistor gain bandwidth product is given by [7]

$$f_T = g_m / 2\pi(C_{be} + C_{jc}) \quad (1)$$

where  $g_m = qI_c/kT$  is the transconductance,  $I_c$  is the quiescent collector current,  $C_{be} = C_{de} + C_{je}$  is the base-emitter capacitance including both junction capacitance  $C_{je}$  and minority carrier diffusion capacitance  $C_{de}$  (which is proportional to  $I_c$ ), and  $C_{jc}$  is the collector junction capacitance. As illustrated in Fig. 2, suppose that required circuit performance demands a transistor gain bandwidth product  $f_T = 120$  MHz which can be satisfied by device A at a collector current  $I_{CA} = 0.20$  mA or by device B at a collector current  $I_{CB} = 6.0$  mA. The choice of device A for low power design is clear. Moreover, for low current operation of both devices

$$f_T \cong (1/2\pi)(qI_c/kT)/(C_{je} + C_{jc}) \quad (2)$$

is directly proportional to  $I_c$  thus indicating the clear advantage of maximizing gain-bandwidth product per unit of quiescent current drain in all transistors used in ana-

log information processing functions. For example, this concept applies in the design of RF receiver circuits for pocket telephones. It clearly suggests a receiver architecture that minimizes use of high frequency front end analog electronics.

A fourth generic principle of low power design that was clearly articulated in antiquity is the advantage of using "extra" electronics to reduce total power drain [7]. This tradeoff of silicon hardware for battery hardware was demonstrated, e.g., for a multi-stage wideband amplifier in which total current drain was reduced by more than an order of magnitude by doubling the number of stages from two to four while maintaining a constant overall gain-bandwidth product [7]. This concept is rather analogous to the approach of scaling down the supply voltage of a CMOS subsystem to reduce its power drain and speed and then adding duplicate parallel processing hardware to restore the throughput capability of the subsystem at an overall savings in power drain [11], [12].

A final overarching principle of low power design that was rigorously illustrated for a wide variety of circuit functions including dc, audio, video, tuned, and low noise amplifiers, nonlinear mixers and detectors and harmonic oscillators as well as bipolar and field effect transistor digital circuits is that micropower design begins with a judicious specification of the required system performance and proceeds to the optimal implementation that fulfills the required performance at minimum power drain [7].

The advent of CMOS digital technology removed quiescent power drain as an unacceptable penalty for broadscale utilization of digital techniques in portable battery operated equipment. Since the average energy dissipation per switching cycle of a CMOS circuit is given by  $E = CV^2$  where  $C$  is the load capacitance and  $V$  is the voltage swing, the obvious path to minimum power dissipation is to reduce  $C$  by scaling down minimum feature size and especially to reduce  $V$ . The minimum allowable value of supply voltage  $V$  for a static CMOS inverter circuit was derived by R. Swanson and the author in 1972 [14] as

$$V_{smin} \geq \beta kT/q \quad (3)$$

where  $\beta$  typically is between 2 and 4.

Early experimental evidence [14] supporting this rigorous derivation is illustrated in Fig. 3, which is a graph of the static transfer characteristic of a CMOS inverter for supply voltages as small as  $V_s = 0.10$  V and matched MOSFET's whose threshold voltages of  $V_t \cong \pm 0.16$  V were controlled by ion implantation. Further discussion of this topic is presented in Section III of this article.

During the 1970's a variety of new micropower techniques were introduced [15], [16] and by far the most widely used product exploiting these techniques was and is the electronic wristwatch [15]. A striking early application of power management occurred in implantable telemetry systems for biomedical research. It was the use of a 15  $\mu$ W 500 kHz monolithic micropower command receiver as an RF controlled switch to connect/disconnect a single 1.35 V mercury cell power source to/from an

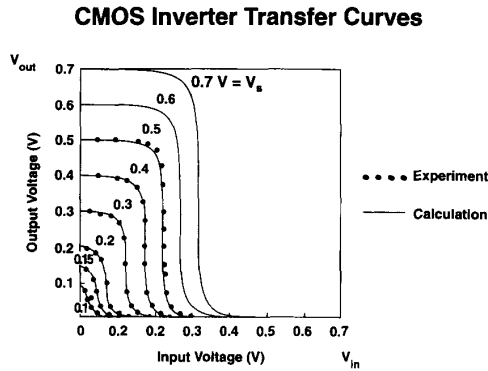


Fig. 3. Static CMOS inverter transfer characteristic [14].

implantable biomedical telemetry system. The fabrication processes used to produce the receiver chip were optimized to yield high value diffused silicon resistors [17]. Entire implantable units including active and passive sensors for biopotential, dimension, blood pressure and flow, chemical ion concentrations, temperature and strain were designed and implemented with power conservation as the primary criterion for optimization [16]. In many respects, the overall system operation of an implantable telemetry unit and its desk-side external electronics subsystem for data processing, display and storage as illustrated in Fig. 4 is similar to but much smaller in scale than the operation of a modern PDA and its backbone network [18].

In the 1980's, the increasing level of power dissipation in mainstream microprocessor, memory and a host of application specific integrated circuit chips prompted an industry wide shift from NMOS and NPN bipolar technologies to CMOS in order to alleviate heat removal problems. The greatly reduced average power drain of CMOS chips provided a relatively effortless interim solution to the problems of low power design. However, the relentless march of microelectronics to higher packing densities and larger clock frequencies has, during the early 1990's, brought low power design to the forefront as a primary requirement for mainstream microelectronics which is addressed in the remainder of this paper.

### III. THEORETICAL LIMITS

#### A. Fundamental Limits

The three most important fundamental limits on low power GSI are derived from the basic physical principles of thermodynamics, quantum mechanics and electromagnetics [19]. Consider first the limit from thermodynamics. Suppose that the node  $N$  illustrated in Fig. 5 is imbedded in a complex microprocessor chip and that between  $N$  and ground  $G$ , there is an equivalent resistance of value  $R$ . Immediately from statistical thermodynamics, it can be shown that the mean square open circuit noise voltage across  $R$  is given by [20]

$$\bar{e}_n^2 = 4kTRB \quad (4)$$

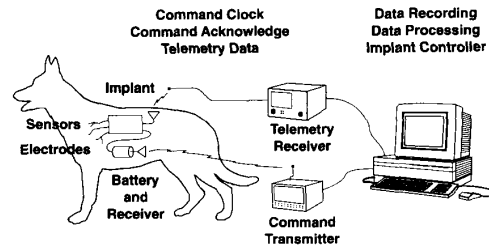


Fig. 4. Implantable telemetry system [16].

#### A Fundamental Limit from Thermodynamics

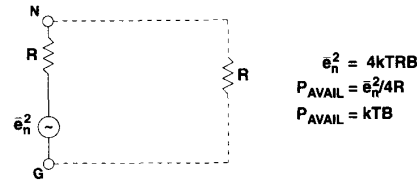


Fig. 5. Model for derivation of fundamental limit from thermodynamics.

and consequently the available noise power is

$$P_{avail} = kTB \quad (5)$$

where  $k$  is Boltzmann's constant,  $T$  is absolute temperature and  $B$  is the bandwidth of the node. Now, it is reasonable to assert that if the information represented at the node is to be changed from a zero to a one, or vice versa, then the average signal power  $P_s$  transfer during the switching transition should be greater than (or at least equal to) the available noise power  $P_{avail}$  by a factor  $\gamma \geq 1$  or

$$P_s \geq \gamma P_{avail}. \quad (6)$$

One can then derive an expression for the switching energy  $E_s$  transfer in the transition,

$$E_s \geq \gamma kT. \quad (7)$$

Clearly, Boltzmann's constant  $k$  and absolute temperature  $T$  are independent of any materials, devices or circuits associated with the node. Consequently,  $E_s$  represents a fundamental limit on binary switching energy. For reasons to be cited in the discussion of circuit limits,  $\gamma = 4$  will be assumed at this point so that at  $T = 300$  K,  $E_s \geq 1.66 \times 10^{-20}$  J = 0.104 eV. One can interpret this limit as the energy required to move a single electron through a potential difference of 0.104 V, which is a Lilliputian energy expenditure compared with current practice which involves energies greater by a factor of about  $10^7$ . One advantage of larger switching energies is that the probability of error due to internal thermal noise energy  $E_n$ ,  $Pr(E_n > E_s)$ , described by a Boltzmann probability distribution function,

$$Pr(E_n > E_s) = \exp(-E_s/kT) \quad (8)$$

decreases exponentially as  $E_s/kT$  increases.

The second fundamental limit on low power GSI is derived from quantum mechanics and more specifically from the Heisenberg uncertainty principle [21], which can be interpreted as follows. A physical measurement associated with a switching transition that is performed in a time  $\Delta t$  must invoke an energy

$$\Delta E \geq h/\Delta t \quad (9)$$

where  $h$  is Planck's constant [19]. Consequently, one can show that

$$P \geq h/(\Delta t)^2 \quad (10)$$

is the required average power transfer during a switching transition of a single electron wave packet. Fig. 6 illustrates the fundamental limits from thermodynamics and quantum mechanics in the power-delay plane. Switching transitions to the left of their loci are forbidden, regardless of the materials, devices or circuits used for implementation. The zone of opportunity for low power GSI lies to the right of these limits. As discussed in Section IV, the power treated in Fig. 6 is a rate of energy transfer and not necessarily a rate of energy dissipation, although the later has virtually always been the case in past practice.

The fundamental limit based on electromagnetics simply dictates that the velocity of propagation  $v$  of a high speed pulse on a global interconnect must be less than the speed of light in free space  $c_0$  or

$$v = L/\tau \leq c_0 \quad (11)$$

where  $L$  is interconnect length and  $\tau$  is interconnect transit time. As illustrated in Fig. 7, the speed of light limit prohibits operation to the left of the  $L = c_0\tau$  locus for any interconnect regardless of the materials or structure used for its implementation.

### B. Material Limits

At the second level of the hierarchy, material limits are independent of the macroscopic geometrical configuration and dimensions of particular device structures. The principal material of interest is Si, which is compared here with GaAs. The primary properties of a semiconductor which determine its key material limits are 1) carrier mobility  $\mu$ , 2) carrier saturation velocity  $v_s$ , 3) self-ionizing electric field strength  $\mathcal{E}_c$ , and 4) thermal conductivity  $K$ . In order to calculate a semiconductor material limit that is independent of the macroscopic properties of a specific device, consider a cube of undoped Si of dimension  $\Delta x$  that is imbedded in a three-dimensional matrix of similar cubes. The material limit on switching energy ( $E = Pt_d$ ) can be calculated as the amount of electrostatic energy stored in this cube of Si of dimension  $\Delta x = V_0/\mathcal{E}_c$  with a voltage difference  $V_0$  across two of its parallel faces, created by an electric field nearly equal to the self-ionizing value  $\mathcal{E}_c$ . Thus

$$Pt_d = E = \epsilon_{Si} V_0^3 / 2\mathcal{E}_c \quad (12)$$

where  $\epsilon_{Si}$  is the permittivity of Si.

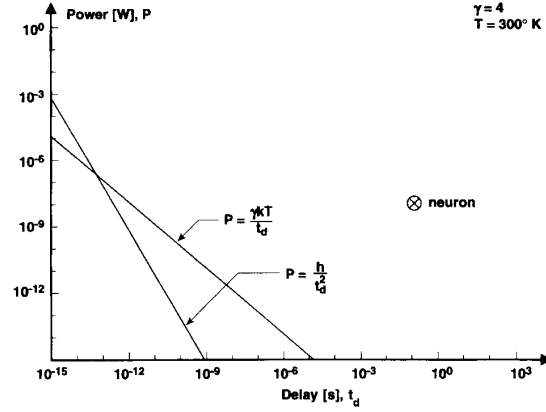


Fig. 6. Average power transfer  $P$  during a switching transition versus transition interval  $t_d$  for fundamental limits derived from thermodynamics and quantum mechanics.

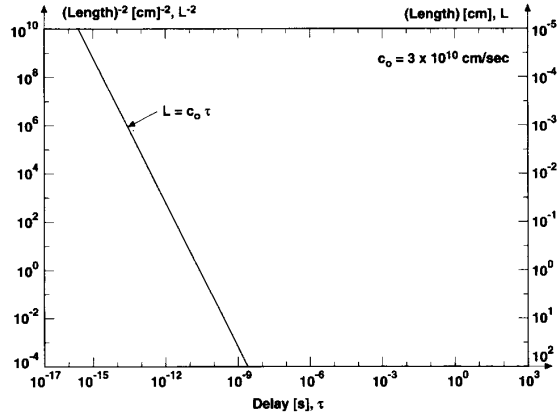


Fig. 7. Square of reciprocal length  $(1/L)^2$  of an interconnect versus interconnect circuit response time  $\tau$  for the fundamental limit from electromagnetics.

The minimum switching time  $t_d$  for this stored energy is taken as the transit time of a carrier through the cube, that is

$$t_d \geq V_0/v_s \mathcal{E}_c \quad (13)$$

describes the material transit time limit. For  $V_0 = 1.0$  V and  $v_s = 10^7$  cm/s,  $t_d = 0.33$  ps for Si and 0.25 ps for GaAs. Thus Si bears only a 33% larger electron transit time per unit of potential drop than GaAs for large values of electric field strength typical for GSI. This small disadvantage is a consequence of two factors: The nearly equal saturation velocities of electrons in Si and GaAs as shown in Fig. 8 as well as the 33% larger breakdown field strength  $\mathcal{E}_c$  of GaAs. Fig. 8 also illustrates the nearly six-fold advantage in electron velocity and therefore mobility that GaAs enjoys at small values of electric field strength (e.g.,  $< 500$  V/cm). In previous generations of technology operating at small values of  $\mathcal{E}$ , it was carrier mobility rather than saturation velocity at large values of  $\mathcal{E}$  (e.g.,  $> 50\,000$  V/cm), that was

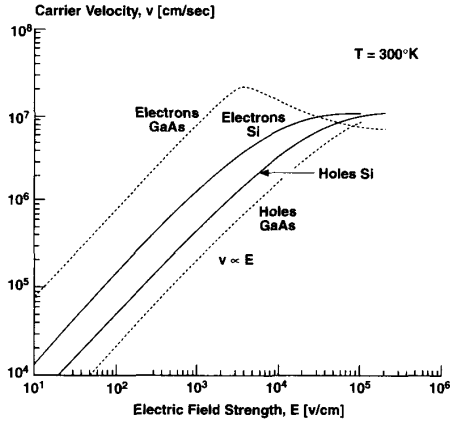


Fig. 8. Carrier velocity versus electric field strength  $\mathcal{E}$  for electrons and holes in Si and GaAs.

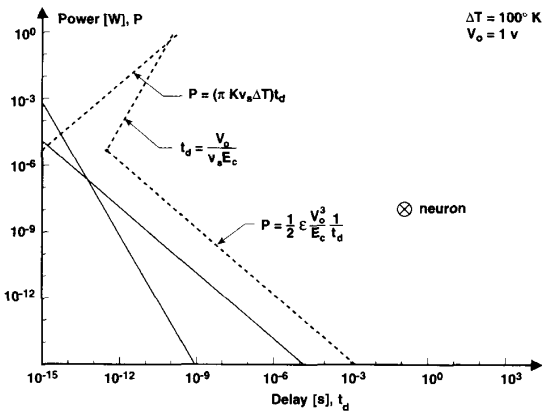


Fig. 9.  $P$  versus  $t_d$  for fundamental limits and Si material limits based on energy storage, transit time, and heat removal.

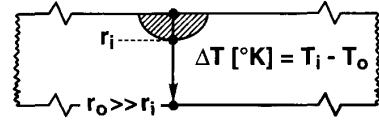
the principal determinate of high speed capability, which is no longer the case.

The switching energy limit for Si given by (12) is illustrated in Fig. 9 for a potential swing  $V_0 = 1.0$  V, which is presumed to be a minimum acceptable value. Solving (13) for  $V_0$  and substituting into (12) gives the locus of minimum switching times

$$P = \epsilon_{Si} \mathcal{E}_c^2 v_s^3 t_d^2 / 2 \quad (14)$$

designated by (13) i.e.,  $t_d \geq V_0 / v_s \mathcal{E}_c$  in Fig. 9. As supply voltage is scaled below approximately 1.0 V, the self-ionization effects that determine  $\mathcal{E}_c$  no longer persist and the limits described by (12)–(14) no longer apply.

In order to derive the heat removal limit at the material level, Fig. 10 illustrates an isolated generic device which is hemispherical in shape with a radius  $r_i$  and located in a chip that is mounted on an ideal heat sink at a temperature  $T_0$ . Based on Fourier's law of heat conduction,  $Q = -KA dT/dx$  where  $Q$  is the heat flow in J/s through an area  $A$  in the presence of a thermal gradient  $dT/dx$ , the power conducted away from the device of diameter



$$\frac{t_s}{P_s} = \frac{1}{\pi K v_s \Delta T}$$

Fig. 10. Model for derivation of material limit based on heat conduction.

$2r_i = v_s t_d$  to the heat sink is given by

$$P = \pi K v_s \Delta T t_d \quad (15)$$

where  $K$  is the semiconductor thermal conductivity and  $T$  is the temperature difference between the device and the heat sink. Substituting representative values indicates that  $t_d/P = 0.21$  ns/W for Si and 0.69 ns/W for GaAs for  $\Delta T = 100$  K. This sample calculation indicates that GaAs suffers a switching time per unit of heat removal that is over 300% greater than the corresponding value for Si, when switching time is limited by substrate thermal conductivity, which is about three times larger for Si than GaAs.

If the device illustrated in Fig. 10 is surrounded by a hemispherical shell of  $\text{SiO}_2$  of radius  $r_s$  representing an SOI structure, the equivalent thermal conductivity  $K_{EQ}$  of the composite structure is given by [22]

$$K_{EQ} = (K_{ox} K_{Si} r_s / r_i) \{ K_{Si} [(r_s / r_i) - 1] + K_{ox} \}^{-1}. \quad (16)$$

Note that as  $r_i \rightarrow r_s$ ,  $K_{EQ} \rightarrow K_{Si}$  and as  $K_{ox} \rightarrow K_{Si}$ ,  $K_{EQ} \rightarrow K_{Si}$ . For  $K_{ox} \cong 0.01 K_{Si}$  and  $r_s = 1.5, 2, 4r_i$ ,  $K_{EQ} \cong 0.029, 0.02, 0.013 K_{Si}$  which indicates a severe reduction in equivalent thermal conductivity of the SOI structure relative to bulk Si.

In summary, Fig. 9 illustrates a second forbidden zone of operation imposed by the characteristics of Si as a material. Operation to the left of the loci of the three material limits defined by (12), (13), and (15) is proscribed for any Si device whose carriers undergo several scattering events within the active region of the device. That is, the three Si material limits illustrated in Fig. 9 assume that the distance over which bulk carrier transport occurs is greater than several mean free path lengths. For shorter distances, the possibility of quasi-ballistic or velocity overshoot effects is best treated in a particular device context [23], [24].

The primary interconnect material limit is defined by substituting a polymer, with a relative dielectric constant  $\epsilon_r \cong 2$ , as the insulator replacing free space in the fundamental speed of light limit, as illustrated in Fig. 11 for the  $L = v\tau$  locus. In essence, both the fundamental and material limits assume a uniform lossless transmission line in a homogeneous dielectric.

### C. Device Limits

Device limits are independent of the particular circuit configuration in which a transistor or an interconnect is applied. The most important device in modern microelectronics is the MOSFET [25] and the most critical limit

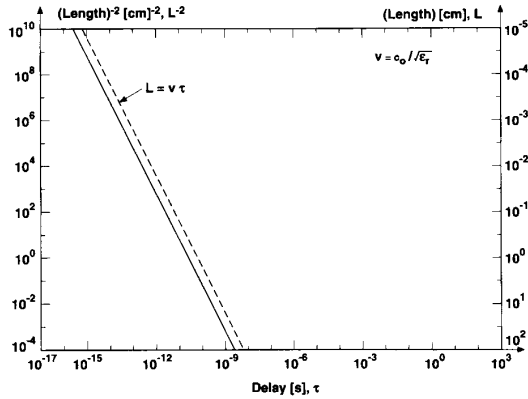


Fig. 11.  $(1/L)^2$  versus  $\tau$  for fundamental limit and material limit with polymer dielectric replacing free space. Both limits are based on the velocity of electromagnetic waves.

on it is its allowable minimum effective channel length  $L_{\min}$  [26]. Consider a family of MOSFET's in which all parameters are held constant except effective channel length  $L$ , which is allowed to take on a wide range of values, e.g.,  $3.0 \mu\text{m} \geq L \geq 0.03 \mu\text{m}$ . As  $L$  is reduced within this range, eventually so-called short channel effects, are manifest [27]. The source and drain depletion regions begin to capture ion charge in the central region of the channel that is strictly under gate control for longer channels. The salient result of such short channel effects which are aggravated as drain voltage increases [28], is that the threshold voltage  $V_t$  is reduced, subthreshold leakage current increases and the MOSFET no longer operates effectively as a switch.

Let us now consider the principal factors which determine the minimum effective channel length  $L_{\min}$  of a MOSFET. In order to achieve  $L_{\min}$ , both gate oxide thickness ( $T_{ox}$ ) and source/drain junction depth ( $X_j$ ) should be as small as possible [29], [30]. Gate leakage currents due to tunneling limit  $T_{ox}$  [31] and parasitic source/drain resistance limits  $X_j$  [32]. In addition, low impurity channels with abrupt retrograde doping profiles are highly desirable for control of short channel effects and high transconductance in bulk MOSFET's [33]. The use of dual gates on opposite sides of a channel ostensibly provides the ultimate structure to contain short channel effects [34]–[36]. Fig. 12 illustrates six different MOSFET structures that have been analyzed consistently to determine their short channel behavior [29] for a very aggressive set of parameters including  $T_{ox} = 3.0 \text{ nm}$  for all devices;  $X_j = 5.0 \text{ nm}$  for shallow junctions;  $X_j = 50 \text{ nm}$ ,  $100 \text{ nm}$  for deep junctions; silicon layer thickness,  $d = 5.5 \text{ nm}$  for the SOI single gate device; and silicon channel thickness,  $d = 10.9 \text{ nm}$  for the dual gate device. Based on both analytical solutions and numerical solutions, using PISCES and DAVINCI, of the two and three dimensional Poisson equation, the threshold voltage roll-off due to scaling of effective channel length is illustrated in Fig. 13 for each of the MOSFET structures sketched in Fig. 12. Families of curves such as these support the prospect of achieving shallow junction

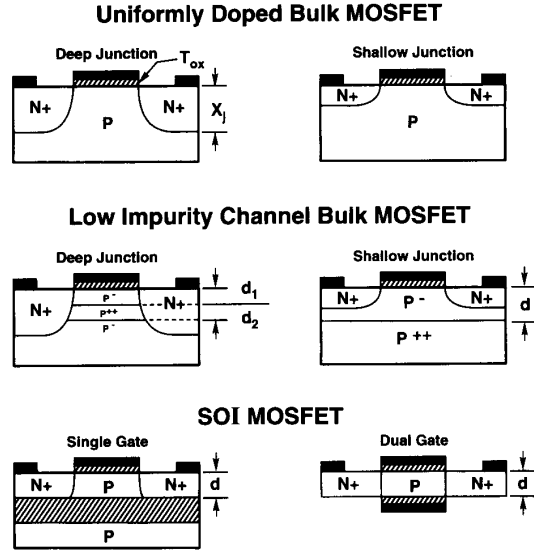


Fig. 12. MOSFET structures.

retrograde channel profile bulk MOSFET's with channel lengths as short as  $60 \text{ nm}$  [29], [30], [32] and dual gate or DELTA MOSFET's with channel lengths as short as  $30 \text{ nm}$  [29], [30], [35]. An interesting feature of the analytical formulation of threshold voltage change [29]

$$\Delta V_T \sim \exp\left\{-\left(1/\pi\right)\left(\epsilon_{ox}/\epsilon_{Si}\right)\left(L/T_{ox}\right)\right\} \quad (17)$$

specifically for the case of deep junctions and uniform doping, but also suggestive of other cases [29], [30] indicates the importance of thin gate insulators with high permittivity in the reduction of short channel effects. In addition to threshold voltage shift, other typically more manageable factors such as bulk punchthrough, gate induced drain leakage and impact ionization which contribute to leakage current and the totality of effects that impact reliability must be observed as potential MOSFET limits [31].

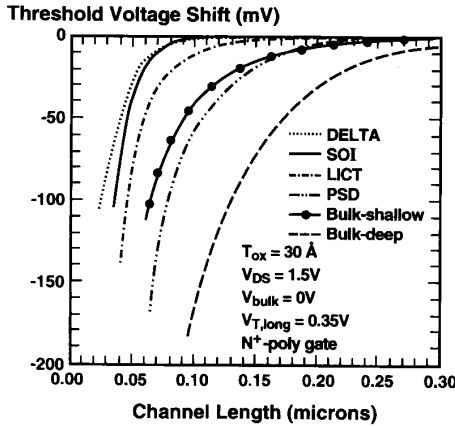
Assuming that analysis of the internal physics of the MOSFET serves to define a minimum effective channel length  $L_{\min}$ , the next stage of effort to define a switching energy limit is to recognize that the relevant energy  $E$  is stored on the gate of the MOSFET at the outset of a switching transition. Therefore, given an allowable minimum effective channel length  $L_{\min}$ , the switching energy limit for a MOSFET is simply

$$E = Pt_d = \frac{1}{2}C_0L_{\min}^2V_0^2 \quad (18)$$

where  $C_0$  is the gate oxide capacitance per unit area corresponding to  $L_{\min}$ . The smallest possible value of the transition time is the channel transit time

$$t_{d\min} = L_{\min}/v_{sc} \quad (19)$$

where  $v_{sc}$  is the saturation velocity of carriers in the channel taken as  $8 \times 10^6 \text{ cm/s}$  for electrons [37]. Assuming 1) minimum feature size  $F \cong L_{\min}$ , 2)  $C_0 = \epsilon_{ox}/T_{ox} =$



**Fig. 13.** Short channel threshold voltage versus channel length for the Si devices shown in Fig. 12. Device parameters at 300 K are: 1) Deep-junction bulk MOSFET:  $N_A = 8.6 \times 10^{17} \text{ cm}^{-3}$ , Junction depth = 1000 Å [29]. 2) Shallow-junction bulk MOSFET:  $N_A = 8.6 \times 10^{17} \text{ cm}^{-3}$ , Junction depth = 50 Å. 3) PSD or low impurity channel deep junction MOSFET:  $N_A = 10^{16} \text{ cm}^{-3}$ ,  $N_A^{++} = 5 \times 10^{18} \text{ cm}^{-3}$ ,  $N_A^+ = 5 \times 10^{17} \text{ cm}^{-3}$ ,  $d = 218 \text{ Å}$ , Junction depth = 500 Å [29]. 4) LICT or low impurity channel shallow junction MOSFET:  $N_A = 5 \times 10^{16} \text{ cm}^{-3}$ ,  $N_A = 5 \times 10^{18} \text{ cm}^{-3}$ ,  $d = 318 \text{ Å}$ , Junction depth = 50 Å. 5) SOI single gate MOSFET:  $N_A = 5 \times 10^{18} \text{ cm}^{-3}$ ,  $d = 55 \text{ Å}$ . 6) Delta or SOI dual gate MOSFET:  $N_A = 5 \times 10^{18} \text{ cm}^{-3}$ ,  $d = 109 \text{ Å}$ .

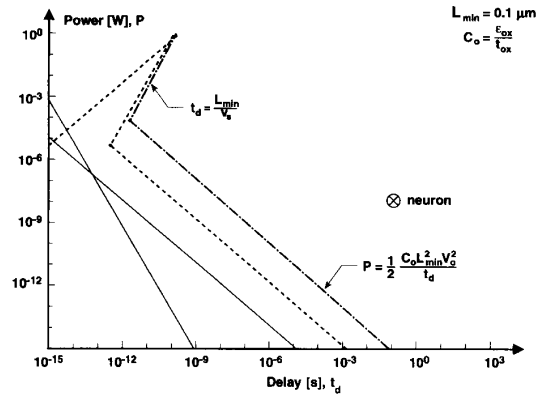
$\epsilon_{ox}/S_{ox}F$  where  $S_{ox}$  is taken as a constant factor relating gate oxide thickness  $T_{ox}$  and minimum feature size  $F$ , and 3)  $V_0 = S_v F$  where  $S_v$  is taken as a constant electric field strength relating supply voltage  $V_0$  and  $F$ , (18) and (19) are solved simultaneously in order to derive the locus of minimum transition times

$$P = \frac{1}{2}(\epsilon_{ox}/S_{ox})S_v^2 v_{sc}^3 t_d^2 \quad \text{or} \quad P = \frac{1}{2}(C_0 L_{min})(V_0/L_{min})^2 v_{sc}^3 t_d^2 \quad (20)$$

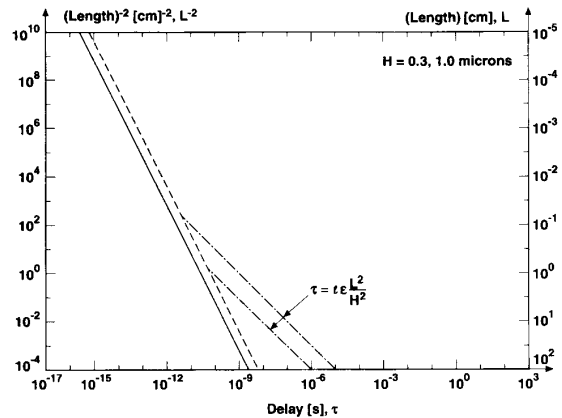
which is designated  $t_d = L_{min}/v_{sc}$  in Fig. 14. The MOSFET switching energy limit (18) and the locus of transition time limits are illustrated in Fig. 14, which includes a third forbidden zone of operation to the left of these loci, for all MOSFET's with channel lengths larger than the conservative value  $L_{min} = 0.1 \mu\text{m}$  and a minimum gate oxide thickness  $T_{ox} = 3.0 \text{ nm}$ . The proximity of the material, (13) and (14), and device, (19) and (20), loci for minimum transition times in Fig. 14 reflects the condition that the electric field strength  $\mathcal{E}$  and carrier velocity  $v_{sc}$  assumed for the MOSFET's are pressing the material limits of Si. A channel saturation velocity of  $8 \times 10^6 \text{ cm/s}$  at a tangential field strength of 200 000 V/cm in a 60 nm channel is quite likely to be somewhat underestimated and more refined values that consider velocity overshoot are needed [24], [35].

The key device limit on interconnects is represented by the response time of a canonical distributed resistance-capacitance network driven by an ideal voltage source. The response of such a network to a unit step function is given in the complex frequency(s) domain as [38]

$$v_0(s) = 1/s \cosh[sRC]^{1/2} \quad (21)$$



**Fig. 14.**  $P$  versus  $t_d$  for fundamental limits, Si material limits and MOSFET device limits derived from gate energy storage and channel transit time.



**Fig. 15.**  $(1/L)^2$  versus  $\tau$  for fundamental limit, material limit and device limits on interconnects for a polymer-copper technology. Device limits represent the response time of a distributed resistance-capacitance network.

where  $R$  and  $C$  are the total resistance and capacitance respectively and it can be shown that the 0–90% response time is  $\tau = 1.0 \text{ RC}$  [38]. In comparison, for a simple RC lumped element model of the distributed network, the 0–90% response time is  $\tau = 2.3 \text{ RC}$ . Neglecting fringing effects, for an interconnect of length  $L$

$$\tau \cong (\rho/H_\rho)(\epsilon/H_\epsilon)L^2 \quad (22)$$

where  $(\rho/H_\rho)$  is the conductor sheet resistance in  $\Omega/\text{square}$  and  $(\epsilon/H_\epsilon)$  is the sheet capacitance in  $\text{F}/\text{cm}^2$ . Fig. 15 illustrates (22) for equal metal and insulator thicknesses  $H_\rho = H_\epsilon = H = 0.3 \mu\text{m}$  and  $1.0 \mu\text{m}$ . A third forbidden zone is evident. For example, no polymer-copper interconnect with a thickness  $H$  smaller than  $0.3 \mu\text{m}$  can operate to the left of the  $0.3 \mu\text{m}$  locus which represents a contour of constant distributed resistance-capacitance product [39].

Further exploration is needed of MOSFET limits to take account of velocity overshoot and random dopant ion placement as well as other effects and of interconnect limits including, for example, inductance and electromigration.

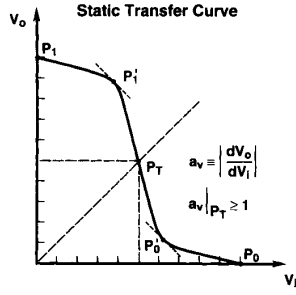


Fig. 16. Static transfer characteristic of a nonideal CMOS inverter.

#### D. Circuit Limits

The proliferation of limits as one ascends the hierarchy necessitates an increasing degree of selectivity in choosing those to be investigated. At the fourth level, circuit limits are independent of the architecture of a particular system. Four key generic circuit limits that cannot be avoided are discussed hereafter. The initial issue to consider is which logic circuit configurations are the most promising for low power microelectronics. The candidates include, e.g., GaAs direct coupled field effect transistor logic (DCFL), Si bipolar transistor emitter coupled logic (ECL), mainstream CMOS and BiCMOS. Of all the logic families now in use, it appears that common static CMOS has the most promise for low power GSI because 1) it has the lowest standby power drain of any logic family, 2) it has the largest operating margins, 3) it is the most scalable, and 4) it is the most flexible in terms of the circuit functions it can implement. For these reasons, this discussion hereafter focuses exclusively on CMOS logic.

The first and foremost generic circuit requirement that must be met by a logic gate is commonly taken for granted. In pursuing limits this practice cannot be followed. It is important to recognize that signal quantization or the capability to distinguish “zeros” from “ones” virtually without error throughout a large digital system is the quintessential requirement of a logic gate. For static CMOS logic this quantization requirement translates into the necessity for an incremental voltage amplification ( $a_v$ ) which is greater than unity in absolute value at the transition point  $P_T$  of the static transfer characteristic of the gate where input and output signals are equal, as illustrated in Fig. 16. This is a heuristic requirement that can be “seen” by considering the need for  $|a_v| > 1$  in order to restore “zero” and “one” levels in an iterative chain of inverters with an arbitrary input level for the initial stage.

An interesting derivative of the requirement for  $|a_v| > 1$  is that the minimum supply voltage  $V_{dd\min}$  for which a CMOS inverter can fulfill the requirement is [14]

$$\begin{aligned} V_{dd} &\geq (2kT/q)[1 + C_{fs}/(C_0 + C_d)] \ln(2 + C_0/C_d) \\ &\geq \beta kT/q \approx 0.1 \text{ V} @ T = 300^\circ\text{K} \end{aligned} \quad (23)$$

where  $\beta$  typically is between 2 and 4 and  $C_{fs}$  is channel fast surface state capacitance,  $C_d$  is channel depletion region

capacitance and  $C_0$  is gate oxide capacitance per unit area in each case. This result which was derived prior to 1972 provides a rationale for selecting a switching energy limit  $E_s = \gamma kT$  at the fundamental level of the hierarchy (7) by postulating that a signal, carried by a single electron charge  $q$  through a potential difference  $\Delta V$ , requires an energy  $q\Delta V = \gamma kT$  and therefore a minimum potential swing  $\Delta V = \gamma kT/q$  defined by (23). Given that the presumed minimum acceptable signal swing for defining both the material (12) and device (18) switching energy limits is  $V_o = 1.0 \text{ V}$ , the question is, “why not set  $V_o = V_{dd\min} = 0.1 \text{ V}$ ?” The simple answer to this question is that to do so would require a threshold voltage  $V_t$  so small that drain leakage current in the off-state of the MOSFET would be entirely too large for most applications. In considering logic and memory circuit behavior at low supply voltages [40]–[48], a value of supply voltage  $V_{dd} = V_o = 1.0 \text{ V}$  appears to be a broadly acceptable compromise between small dynamic power and small static power dissipation, although confirmation of (23) in a low power system with  $V_{dd} = 200 \text{ mV}$  is a prominent recent development [48].

Assuming negligible static power drain due to MOSFET leakage currents, a second generic circuit limit on CMOS technology is the familiar energy dissipation per switching transition

$$E = Pt_d = \frac{1}{2} C_c V_o^2 \quad (24)$$

where  $C_c$  is taken as the total load capacitance of a ring oscillator stage, including output diffusion capacitance, wiring capacitance and input gate capacitance for an inverter which occupies a substrate area of  $100 F^2$  where the minimum feature size  $F = 0.1 \mu\text{m}$ .

Assuming carrier velocity saturation in the MOSFET’s, an approximate value of the drain saturation current is

$$I_{ds} \approx Z C_o v_{sc} (V_g - V_t) \quad (25)$$

where  $Z$  is the channel width,  $V_t$  is threshold voltage and the gate voltage,  $V_g = V_o$ . A third generic circuit limit, the intrinsic gate delay can be calculated as  $t_d = 1/2[C_c V_o / I_{ds}]$  or using (25)

$$t_d = \frac{1}{2} [C_c / Z v_{sc} C_o] [V_o / (V_o - V_t)] \quad (26)$$

assuming that the product  $Z v_{sc} C_o$  is equal for the  $N$  and  $P$ -channel MOSFET’s and that their threshold voltages are matched. For 1)  $C_c \cong S_c F$  where  $S_c$  is taken as a constant factor relating load capacitance  $C_c$  and minimum feature size  $F$ , 2)  $V_o = S_v F$  as in the derivation of (20), 3)  $Z = S_Z F$  where  $S_Z$  is a constant relating channel width  $Z$  and  $F$ , 4)  $C_o = \epsilon_{ox} / S_{ox} F$  as in the derivation of (20), and 5)  $V_o / (V_o - V_t) \cong 1$ , (24) and (26) are solved simultaneously for the locus of intrinsic gate delay times,

$$P = 4(V_o / C_c)^2 (Z C_o v_{sc})^3 t_d^2 \quad (27)$$

which is designated  $t_d \cong \frac{1}{2} C_c / Z C_o v_{sc}$  in Fig. 17. The CMOS circuit switching energy limit (24) and the locus of intrinsic gate delay times (27) are illustrated in Fig. 17 which includes a fourth forbidden zone, to the left of

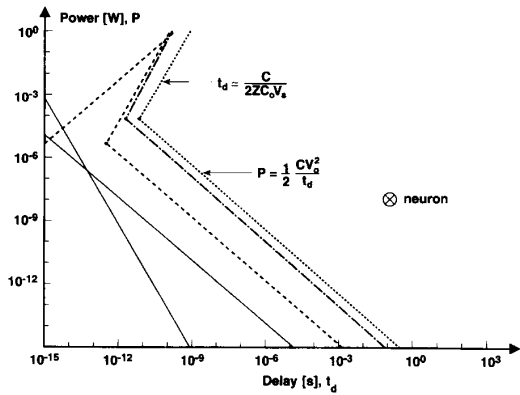


Fig. 17.  $P$  versus  $t_d$  for fundamental, Si material, MOSFET and CMOS circuit limits. Circuit limits are derived from switching energy and intrinsic gate delay analyses.

these loci, for all CMOS circuits using bulk technology with feature sizes larger than  $F = 0.1 \mu\text{m}$ . A gate oxide thickness  $T_{ox} = 3.0 \text{ nm}$  and a load capacitance  $C_c = 0.5 \text{ fF}$  are used in plotting the loci of Fig. 17.

The fourth generic circuit limit applies to a transistor driving a global interconnect presented as a distributed resistance-capacitance network extending, e.g., between opposite corners of a chip. The response time of this global interconnect circuit is [49]

$$\tau \cong (2.3R_{tr} + R_{int})C_{int} \quad (28)$$

where  $R_{tr}$  is the output resistance of the transistor driver and  $R_{int}$  and  $C_{int}$  are the total resistance and capacitance, respectively of the global interconnect. To prevent excessive delay due to wiring resistance, the circuit should be designed so that  $R_{int} < 2.3R_{tr}$  giving

$$\tau \cong 2.3R_{tr}C_{int} = 2.3R_{tr}c_{int}L \quad (29)$$

where  $c_{int}$  is the capacitance per unit length of the interconnect. The distributed capacitance of a nearly lossless or TEM-mode transmission line can be expressed as  $c_{int} = 1/vZ_o$  where  $v = c_o[\epsilon_r]^{-1/2}$  is the wave propagation velocity of the line,  $\epsilon_r$  is the relative permittivity of its dielectric,  $Z_o \approx [\mu_o/\epsilon_o\epsilon_r]^{1/2}$  is its characteristic impedance and  $c_o = 1/[\mu_o\epsilon_o]^{1/2}$ . This global interconnect response time limit is illustrated in Fig. 18. The region to the left of the

$$\tau \cong 2.3R_{tr}C_{int} = 2.3(R_{tr}/Z_o)(L/v) \quad (30)$$

locus is a forbidden zone for driver resistances larger than the designated value for the locus. Although the interconnect models engaged in the preceding discussion are rather elementary, they provide a clear picture of circuit limits which govern global interconnect performance.

Further exploration is needed of limits imposed by other important circuit configurations and by new device structures.

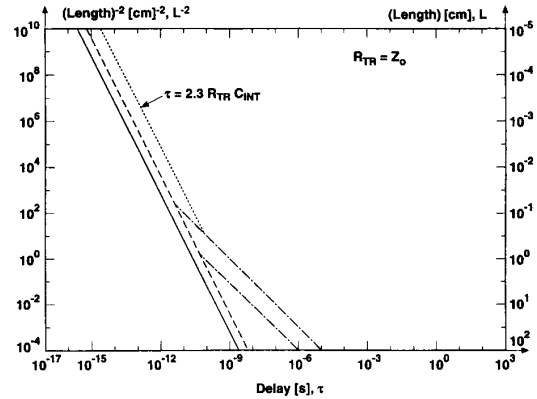


Fig. 18.  $(1/L)^2$  versus  $\tau$  for fundamental, material, device, and circuit limits on interconnects. Circuit limits represent the response time of a circuit consisting of a MOSFET driving a lumped interconnect capacitance.

### E. System Limits

System limits are the most numerous and nebulous ones in the hierarchy. They depend on all other limits and above all they are the most restrictive ones in the hierarchy. Consequently, it is imperative that these predominant limits be carefully considered. Among the innumerable constraints arising from the fact that each different chip design has its own unique set of limits, there are five inescapable generic system limits that are elucidated in this discussion. These limits are set by 1) the architecture of the chip, 2) the power-delay product of the CMOS technology used to implement the chip, 3) the heat removal or cooling capacity of the chip package, 4) the cycle time requirements imposed on the chip, and 5) its physical size. To illustrate these generic limits it is necessary to select a particular example for a case study, which is intended to be broadly applicable. In keeping with the intent to explore opportunities for low power GSI, salient boundary conditions that are assumed for the case study are: 1) a generic architecture equivalent in complexity to one billion logic gates, i.e.,  $N_g T = 10^9$ , 2) CMOS technology with a conservative minimum feature size,  $F = 0.1 \mu\text{m}$ , 3) a package cooling coefficient,  $Q = 50 \text{ W/cm}^2$ , 4) a clock frequency  $f_c = 1.0 \text{ GHz}$ , and 5) a single chip implementation.

A block diagram of the system architecture is illustrated in Fig. 19. It is conceived as a two-dimensional systolic array [50]–[52] of 1024 identical macrocells, each consisting of a number of gates  $N_g = 10^9/1024$ . Communication between macrocells is assumed to occur only at the physical boundaries of adjacent macrocells. Each macrocell is assumed to receive an unskewed clock signal distributed by a balanced five-level H-tree network [38], [53], to the geometric center of the macrocell. Both logic and timing signals are communicated over arbitrary paths within a square macrocell of dimension  $L$  so the maximum path length for clock skew is  $L$  and the maximum logic signal path length is  $2L$ . The macrocell is represented as a random

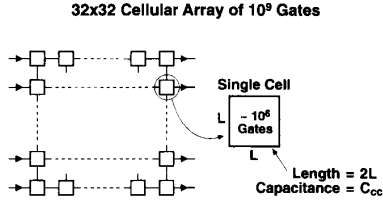


Fig. 19. Systolic array system block diagram.

logic network described by Rent's rule [54]

$$N_p = K_p N_g^p \quad (31)$$

where  $N_p$  is the number of signal lines entering or exiting the macrocell, Rent's coefficient  $K_p = 0.82$  and Rent's exponent  $p = 0.45$  which are empirically determined values for microprocessors [38]. Using Rent's rule as the basis of a stochastic analysis, the average length of an interconnect in gate pitches  $\bar{R}$  can be calculated as [55]

$$\begin{aligned} \bar{R}_{rl} = 2/9 \{ & 7(N_g^{p-0.5} - 1)(4^{p-0.5} - 1)^{-1} \\ & - (1 - N_g^{p-1.5})(1 - 4^{p-1.5})^{-1} \} \\ & \cdot (1 - 4^{p-1})(1 - N_g^{p-1})^{-1} \end{aligned} \quad (32)$$

for  $p \neq 0.5$ . For the microprocessor-like macrocell with  $p = 0.45$ , (32) gives  $\bar{R}_{rl} \cong 6$ . Thus the total wire length loading a gate in the random logic network is

$$l_{rl} = \bar{R}_{rl} \cdot FO \cdot [A_{rl}]^{1/2} \quad (33)$$

where  $FO = 3$  is the fan-out and gate area  $A_{rl} = 200 F^2$  is assumed to be limited by transistor packing density. This places a stringent demand on local wiring area which requires a logic gate dimension [38]

$$[A_{rl}]^{1/2} = \bar{R}_{rl} \cdot FO \cdot p_w / e_w n_w \quad (34)$$

where  $n_w$  is the number of wiring levels,  $p_w$  is the wiring pitch and  $e_w$  is the wiring efficiency factor. For  $n_w = 4$ ,  $p_w = 0.2 \mu\text{m}$ , and  $e_w = 0.75$ , as well as for  $n_w = 6$ ,  $p_w = 0.2 \mu\text{m}$ , and  $e_w = 0.5$  transistor packing density limits logic gate area.

As illustrated in Fig. 20 the system switching energy limit is defined by a composite gate which characterizes the critical path of a macrocell. For a logic signal this path is assumed to consist of 1) a chain of  $n_{cp}$  random logic gates and 2) one macrocell corner-to-corner global interconnect of length  $2L$  [56], [57]. Therefore, the prorata switching energy of the composite gate is given by

$$E = P t_d = \frac{1}{2} C_{rl} [1 + C_{cc}/n_{cp} C_{rl}] V_o^2 \quad (35)$$

where  $C_{rl}$  is the total capacitance loading a random logic gate including MOSFET diffusion capacitance, wiring capacitance for a total interconnect length  $l_{rl}$  (33), and MOSFET gate capacitance, and  $C_{cc}$  is the total capacitance of the corner-to-corner interconnect circuit. The effective propagation delay time of the composite gate is defined as

$$t_d = t_{drl} (1 + T_{cc}/n_{cp} t_{drl}) \quad (36)$$

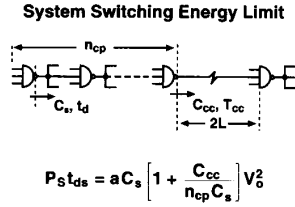


Fig. 20. Critical path used to define the system switching energy limit.

where  $t_{drl}$  is the delay time of a random logic gate and  $T_{cc}$  is the response time of the corner-to-corner interconnect circuit. In (35),  $P$  is the average power dissipation of a composite gate during the propagation delay time  $t_d$ .

As illustrated in Fig. 21, the system heat removal limit is defined by the requirement that the average power dissipation of a composite gate  $\bar{P}$  must be less than the cooling capacity of the packaging or

$$\bar{P} = aE/T_c \leq QA \quad (37)$$

where  $E$  as in (35) is the switching energy of a composite gate,  $a \leq 1$  is the probability that the gate switches during a clock cycle,  $T_c = 1/f_c$  is the clock period,  $Q$  [W/cm<sup>2</sup>] is the package cooling coefficient and  $A$  is the substrate area occupied by the critical path composite gate. It is assumed that the composite gate area  $A$  consists of the area occupied by a random logic gate  $A_{rl}$  plus a *pro rata* share of the area of the corner-to-corner driver circuit  $A_{cc}$  and that  $A_{rl}/A_{cc} = C_{rl}/C_{cc}$  which gives

$$A = A_{rl} (1 + C_{cc}/n_{cp} C_{rl}). \quad (38)$$

The cycle time can be expressed as

$$T_c = s_{cp} n_{cp} t_d \quad (39)$$

where it is to be shown that  $s_{cp} \geq 1$  accounts for a small clock skew [58]. Combining (35), (37)–(39)

$$P \leq (s_{cp} n_{cp}/a) Q A_{rl} (1 + C_{cc}/n_{cp} C_{rl}) \quad (40)$$

gives the maximum allowable value of  $P$ , that is permitted by the cooling capability of the package and therefore the minimum composite gate delay  $t_d$  as defined by (35). Assuming 1)  $C_{rl} \cong S_{rl} F$  and  $C_{cc} = S_{cc} F$  where  $S_{rl}$  and  $S_{cc}$  are constants relating, respectively, random logic gate load capacitance  $C_{rl}$  and corner-to-corner interconnect capacitance  $C_{cc}$  to minimum feature size  $F$ , 2)  $V_o = S_v F$  as in the derivation of (20), and 3)  $A_{rl} = S_{rl} F^2$  where  $S_{rl}$  is a constant relating random logic gate area  $A_{rl}$  to  $F^2$ , and solving (35) and (40) simultaneously gives the locus of minimum composite gate delays

$$P = (1 + C_{cc}/n_{cp} C_{rl}) (\frac{1}{2} C_{rl} V_o^2)^{-2} [(s_{cp} n_{cp}/a) Q A_{rl}]^3 t_d^2 \quad (41)$$

which is designated as  $\bar{P} \leq QA$  in Fig. 22. In plotting (41), in addition to the parameter values listed in the figure,

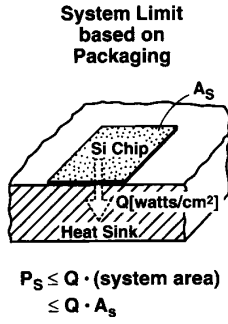


Fig. 21. System heat removal limit based on packaging.

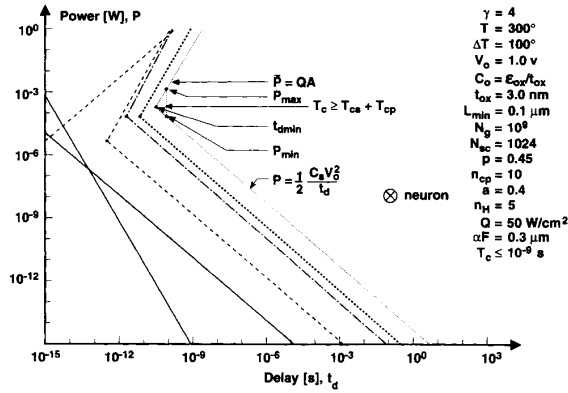


Fig. 22.  $P$  versus  $t_d$  for fundamental, material, device, circuit, and system limits. System limits are imposed by switching energy ( $Pt_d = \frac{1}{2}C_s V_o^2$ ), heat removal ( $\bar{P} \leq QA$ ), and cycle time ( $T_c \geq T_{cs} + T_{cp}$ ) requirements.

$C_{cc} = 100$  fF,  $C_{rl} = 3.28$  fF,  $A_{rl} = 2 \times 10^{-8}$  cm<sup>2</sup> and  $s_{cp} = 1.11$ .

The system cycle time limit is given by

$$T_c \geq T_{cs} + T_{cp} \quad (42)$$

where  $T_{cs}$  is the maximum clock skew within a macrocell,  $T_{cp} = n_{cp}t_d$  is the critical path delay and  $t_d$  is given by (36). From (42)

$$t_{dmax} \leq (T_c - T_{cs})/n_{cp} \quad (43)$$

is the maximum allowable value of critical path composite gate delay that enables the required cycle time. If the allowed clock skew is  $T_{cs} = \eta T_c$  (e.g.,  $\eta \approx 0.1$ ) then referring to (39),  $s_{cp} = 1/(1 - \eta) \approx 1.11$ . To calculate  $t_d$  as given by (36) at the system level, appropriate values are used in (26) to compute the random logic gate delay  $t_{drl}$  and in (28) for both logic and clock global interconnects assuming  $2.3R_{tr} = R_{int}$  and, referring to (22),  $H_\rho = H_\epsilon = 0.3 \mu\text{m}$ . In calculating both  $T_{cs}$  and  $T_{cc}$  the macrocell size is taken as  $L^2 \cong N_g A_{rl}$ .

The system critical path composite gate switching energy, heat removal and timing limits are illustrated in Fig. 22. Operation to the left of the switching energy

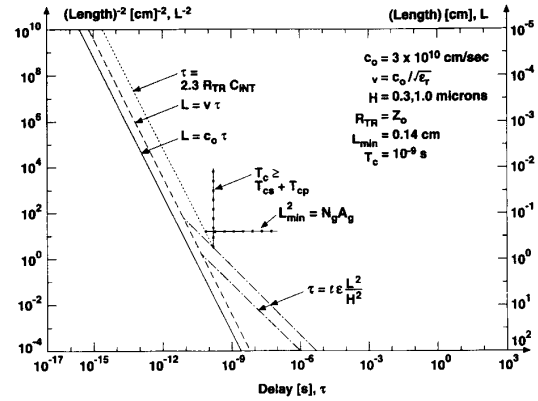


Fig. 23.  $(1/L)^2$  versus  $\tau$  for all levels of the hierarchy. System limits are imposed by global interconnect response time designated by  $T_c \geq T_{cs} + T_{cp}$  and length designated by  $L_{min}^2 = N_g A_g$ .

locus designated  $P_s t_d = 1/2 C_s V_o^2$  and the heat removal locus designated  $\bar{P} \leq QA$  is forbidden and operation to the right of the timing locus  $T_c \geq T_{cs} + T_{cp}$  is also forbidden. The allowable design space for this particular macrocell is the small triangle with vertices 1)  $t_{dmin}$  corresponding to minimum achievable propagation delay for the composite gate and therefore to the maximum performance design, 2)  $P_{min}$  corresponding to the lowest composite gate switching power that provides the required clock frequency  $f_c = 1.0$  GHz and 3)  $P_{max}$  corresponding to the most mature technology or largest minimum feature size and chip size that provides  $f_c = 1.0$  GHz. The three sides of the triangle correspond to contours of 1) constant switching energy  $E$  reflecting the performance level of the MOSFET and interconnect technologies, 2) constant heat removal capacity  $Q$  reflecting the performance level of the packaging technology, and 3) constant clock period  $T_c$  reflecting the performance level required by the design.

At the system level,  $(1/L)^2$  versus  $\tau$  plane limits focus on the longest interconnects since typically they impose the most stringent demands on performance. As illustrated in Fig. 23, the response time of the longest global interconnect, i.e., a logic signal path, of length  $2L$  is designated by  $T_c \geq T_{cs} + T_{cp}$  since  $T_c \geq T_{cs} + T_{cp} = T_{cs} + T_{cc} + n_{cp}t_{drl}$ , and  $T_{cc} = T_c - T_{cs} - n_{cp}t_{drl}$  is the response time of the longest global interconnect. The actual length of its path is designated  $L_{min}^2 = N_g A_g$ , since a smaller area than  $L_{min}^2$  could not accommodate the required number of logic gates  $N_g$  using the prescribed technology which requires a gate area  $A_g$ . The longest global interconnect cannot have a slower response time nor a smaller length than designated by these two limits. The forbidden zone of operation for the longest interconnects lies external to the small triangle, two of whose sides are defined by the preceding limits. The size of this triangle appears to be almost vanishingly small, particularly as a result of the stringent demands of a 1.0 GHz clock frequency. For smaller values of  $f_c$ , the size of the triangle can be enlarged at the expense of reduced performance.

The distinctive feature of the preceding treatment of system limits is that it seeks to describe the unbounded range of options of system architecture (not to mention algorithms) in terms of the absolute minimum number of parameters that enable a concise definition of the generic physical limits on system performance and hence a revealing juxtaposition of these system limits with the full hierarchy of limits which governs opportunities for GSI. A salient feature of the system representation is the definition of a critical path and from that the derivation of a composite logic gate which performs canonical computational operations. Only the first rudimentary results of this approach to low power system simulation, as illustrated in Figs. 22 and 23 are available at this juncture.

#### IV. QUASI-ADIABATIC MICROELECTRONICS

During an adiabatic process no loss or gain of heat occurs. A quasi-adiabatic process is designed to resemble this ideal behavior. The fundamental opportunity of quasi-adiabatic microelectronics is based on the second law of thermodynamics, which can be stated as follows: In any thermodynamic process that proceeds from one equilibrium state to another, the entropy of a closed system either remains unchanged or increases [59]. Entropy change  $dS$  can be expressed as

$$dQ/T = dS \geq 0 \quad (44)$$

where  $dQ$  is the heat added to the system and  $T$  is its absolute temperature. In a computational process, it is only those steps that discard information or increase disorder and therefore increase entropy ( $dS > 0$ ) which have a lower limit on energy dissipation or heat generation ( $dQ > 0$ ) imposed by the second law of thermodynamics [4], [5]. Consequently, the intriguing prospect of inventing quasi-adiabatic computational technology offers the possibility of reducing power dissipation to levels below those imposed by limits on the nonadiabatic processes discussed in Section III.

To elucidate this principle, consider the circuit operation illustrated in Fig. 24 in which the capacitor  $C$  is charged through the resistor  $R$  from a voltage source  $V_{in}$ . If  $V_{in}$  changes as a step function, the energy dissipated in  $R$  while charging  $C$  to a voltage  $V_o$  is given by

$$E_{d1} = \frac{1}{2} CV_o^2. \quad (45)$$

However, if  $V_{in}$  changes as a very slowly varying ramp function of rise time  $T_d \gg 2RC$ , then the energy dissipated in  $R$  is given by

$$E_{d2} \cong \frac{1}{2} CV_o^2 [2t_d/T_d] \quad (46)$$

where  $t_d = RC$ . Since  $2t_d/T_d \ll 1$ ,  $E_{d2} \ll E_{d1}$ . In fact, one might say that  $E_{d2}$  describes an asymptotically vanishing amount of energy as  $T_d \rightarrow \infty$  [60]. The reduction in energy dissipation is a consequence of maintaining at all times a quasi-equilibrium condition for which  $V_{in} \rightarrow V_{out}$  to keep the current flow nearly zero so that

$$E_d = \int_0^\infty i^2 R dt \rightarrow 0. \quad (47)$$

#### Quasi-Adiabatic Switching

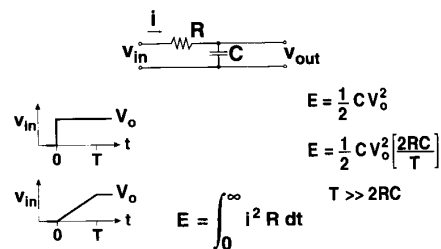


Fig. 24. Quasi-adiabatic switching.

Moreover, the discharge of  $C$  through  $R$  must be achieved through a very slowly varying ramp function whose fall time is  $T_d \gg 2RC$ . And, the source voltage generator providing  $V_{in}$  must include highly efficient resonant circuits to enable recycling a major fraction of the transferred energy.

The two key requirements for quasi-adiabatic or asymptotically zero dissipation digital microelectronics are summarized by (44) and (47): information cannot be destroyed and quasi-equilibrium operation must prevail [59], [60]. These requirements can be reflected in a hierarchy of limits on quasi-adiabatic microelectronics as illustrated in Fig. 25, which graphs the energy dissipation  $E_d$  during a switching transition versus the ratio of external transition time  $T_d$  to twice internal transition time  $2t_d = 2RC$  or  $E_d$  versus  $T_d/2t_d$ . The salient message that Fig. 25 conveys is that, in principle, external control of the switching transition time (e.g., via a slow ramp of supply voltage as illustrated in Fig. 24) causing  $T_d/2t_d \gg 1$  can reduce switching energy dissipation  $E_d$  to arbitrarily small amounts. For logarithmic scales, diagonal lines in the  $E_d$  versus  $T_d/2t_d$  plane represent loci of constant switching energy transfer  $E$ , which is precisely the case for the  $P$  versus  $t_d$  plane. Consequently, it becomes clear that  $P$  versus  $t_d$  plane limits serve as helpful benchmarks in assessing performance of quasi-adiabatic microelectronics. At the fundamental level of the hierarchy  $E_d = \gamma kT$  at  $T_d/2t_d = 1$  corresponds to the fundamental limit from thermodynamics (7) as previously shown in the  $P$  versus  $t_d$  plane by Fig. 6. At the material level,  $E_d = 1/2 \epsilon_{Si} V_o^3 / \epsilon_c$  at  $T_d/2t_d = 1$  as given by (12) and Fig. 9 and for the device level  $E_d = 1/2 C_0 L_{min}^2 V_o^2$  at  $T_d/2t_d = 1$  as given by (18) and Fig. 14. The capability to illustrate fundamental, material and device limits in the  $E_d$  versus  $T_d/2t_d$  plane is predicated on the assumption that the associated switching behavior can be enabled by means that are unspecified. While this assumption serves to add insight at the first three levels of the hierarchy, it should not be casually engaged at the fourth level because unlike the unchanging materials and device structures of the second and third levels, the circuit configurations used for nonadiabatic operation, such as static CMOS, must change very significantly for quasi-adiabatic operation [60]–[62]. Consequently, without identifying specific quasi-adiabatic circuit topologies and system architectures, the circuit and

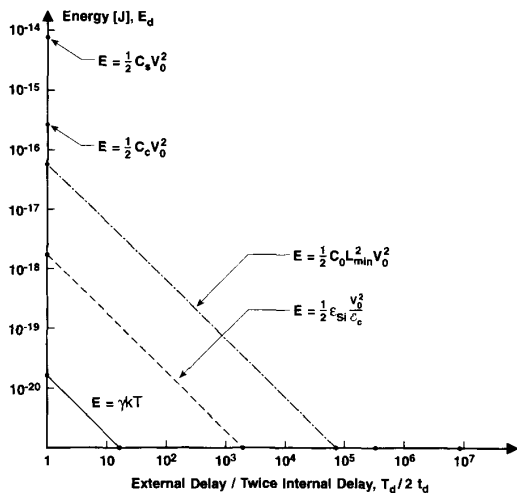


Fig. 25. Energy dissipation  $E_d$  during a switching transition versus the ratio of twice internal transition time  $2t_d$  to external transition time  $T_d$ .

system limits that would be displayed in Fig. 25 must be held in abeyance. Invention is expected to abbreviate the delay in completing the  $E_d$  versus  $T_d/2t_d$  plane hierarchy in which quasi-adiabatic operation will improve on the conventional nonadiabatic circuit and system level benchmarks defined by (24) and (35) respectively and illustrated as single points on the  $T_d/2t_d = 1$  axis in Fig. 25 [62].

The current surge of interest in quasi-adiabatic circuit and system techniques [61] underscores the importance of low power microelectronics and thus the establishment of wholistic approaches to minimization of energy expenditure as exemplified by the hierarchy of limits explored in this discussion.

## V. PRACTICAL LIMITS

In dealing with practical limits, the key question is "How many transistors can we expect to fabricate in a single silicon chip that will prove to be useful at some specific time in the future?" To gain insight into this issue, the number of transistors per chip  $N$  can be elegantly expressed in terms of three macrovariables:  $N = F^{-2} \cdot D^2 \cdot PE$  [1], [3].

The evolution of average minimum feature size  $F$  for state-of-the-art microchips is described in simplified form in Fig. 26, which is a graph of  $F$  versus calendar year  $Y$ . In 1960,  $F$  was about 25  $\mu\text{m}$ . By 1980, it had scaled down to 2.5  $\mu\text{m}$ . If the historical rate of evolution continues throughout the 1990's,  $F$  will be about 0.25  $\mu\text{m}$  in the year 2000. Following that, Fig. 26 illustrates three possible scenarios: 1) the 0.25  $\mu\text{m}$  or pessimistic scenario, 2) the 0.125  $\mu\text{m}$  or realistic scenario, and 3) the 0.0625  $\mu\text{m}$  or optimistic scenario. The pessimistic scenario simply projects no further reduction of  $F$  beyond 0.25  $\mu\text{m}$  based on the adverse expectation that the cost per function or the cost per logic circuit of a microchip will reach a minimum for the design, manufacturing, testing and packaging technologies

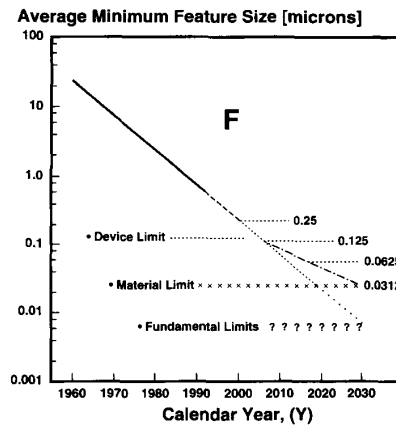


Fig. 26. Average minimum feature size  $F$  versus calendar year  $Y$ .

required by the 0.25  $\mu\text{m}$  generation. This scenario seems unlikely at this time except to pessimists. The realistic scenario projects a further reduction of  $F$  at the historic rate until the later years of the first decade of the next century. Then saturation occurs at 0.125  $\mu\text{m}$  again based on the economic expectation that the cost per function of a microchip will reach a minimum for the 0.125  $\mu\text{m}$  generation especially because it could be the last generation for which deep ultraviolet microlithography will suffice. The optimistic scenario projects further reduction of  $F$  at a slower rate resulting in about 0.0625–0.0500  $\mu\text{m}$  average minimum feature size during the second decade of the millennium, and then saturation. The slower rate of reduction and saturation of  $F$  at 0.0625  $\mu\text{m}$  could be caused by a combination of factors, including astronomical capital costs particularly due to introduction of a radically different microlithography technology (e.g., using X-rays) and a soft collision with the physical limits on dimensions of MOSFET's finally imposing a minimum cost per function on the 0.0625  $\mu\text{m}$  generation. The author's estimate is that CMOS microchips with minimum feature sizes in the 0.0625  $\mu\text{m}$  range will be widely used.

Historically, the advantages of larger chip area have been reduced cost per function, improved performance, enhanced reliability and smaller size and weight at the module, board or box level for microelectronic equipment. The evolution of the square root of microchip area or chip size is illustrated in simplified form in Fig. 27. In 1960,  $D = \sqrt{\text{chip area}}$  was about 1.2 mm; in 1980, about 6.5–7.0 mm; and, if the recent historic rate of increase continues throughout the 1990's,  $D$  will reach a range around 25 mm in the year 2000. Thereafter, three possible scenarios are again illustrated by segments F, G and H. Scenario F pessimistically projects saturation of  $D$  at about 25 mm based on a maximum silicon wafer diameter of 200 mm. A realistic scenario for the 2000–2010 period is that 300 mm wafers will be commonly used and that chip sizes up to 40 mm will be economic. Beyond this a long range optimistic scenario projects 400 mm wafers and over 50 mm chips.

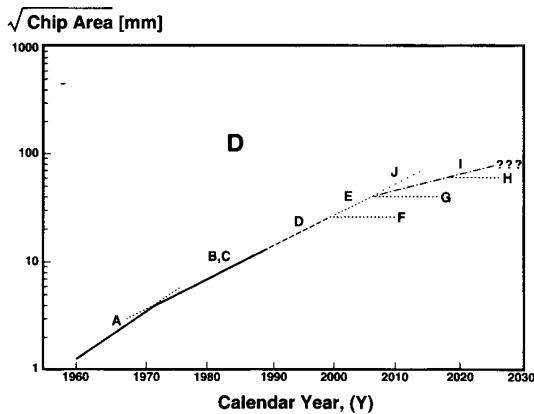


Fig. 27. Square root of die area  $D$  versus calendar year  $Y$ .  $\sqrt{\text{die area}} = D$ .

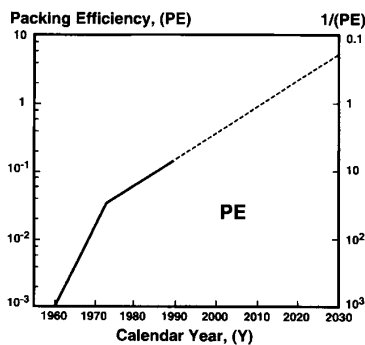


Fig. 28. Packing efficiency  $PE$  versus calendar year  $Y$ . Note that packing efficiency is defined as the number of transistors per minimum feature area.

The third macrovariable that contributes to the growing number of transistors per chip is their packing efficiency  $PE$ , the number of transistors per minimum feature area. The most prominent feature of the evolution of  $PE$ , presented in simplified form in Fig. 28, is the abrupt change in the slope of the locus which occurred in the early 1970's. Its cause was the unavailability of silicon real estate on the chip. Prior to about 1972,  $PE$  was increased simply by moving transistors and metal interconnects closer together. Since 1972, improvements in  $PE$  have been achieved by extending into the third dimension through increasing the number of mask levels in a chip manufacturing sequence. This trend toward clever use of the third dimension is not expected to change. It is interesting that about 2010,  $PE$  approaches unity; that is the areal packing efficiency is projected as one transistor per minimum feature area, which is truly a three-dimensional microchip suggesting multiple levels of transistors.

A simplified composite curve illustrating the number of transistors per chip  $N$  versus calendar year is shown in Fig. 29. This graph more than any other chronicles the progress of the microchip from its inception in 1959, until

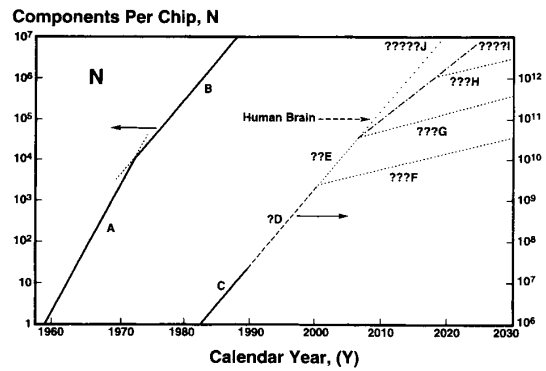


Fig. 29. Number of transistors per chip  $N$  versus calendar year  $Y$ .

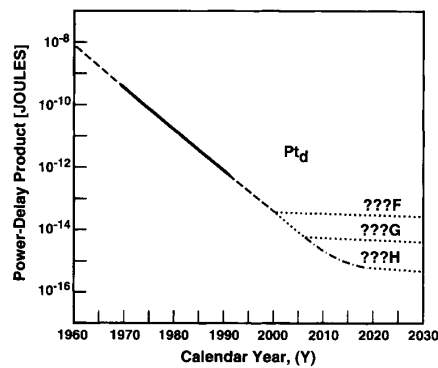


Fig. 30. System level power-delay product  $Pt_d$  versus calendar year  $Y$ .

1995 and beyond. The pessimistic scenario denoted by segment  $F$  projects a one-billion transistor chip or GSI by the year 2000, a forecast first proposed by the author in 1983 [3]. The realistic scenario projects over 100 billion transistors per chip before the year 2020.

One can also graph switching energy or power-delay product ( $Pt_d$ ) versus calendar year as illustrated for CMOS technology in Fig. 30, again for three possible future scenarios [62]. Finally, the chip performance index  $CPI$  can be calculated as the quotient, of  $N$  and  $Pt_d$  or  $CPI = N/Pt_d$ . As illustrated in Fig. 31, the  $CPI$  has grown by about twelve decades since 1960 and is realistically projected to grow by about another six decades before 2020. This enormous rate of both productivity and performance enhancements is unprecedented in technological history.

## VI. CONCLUSION

Historically there can be no doubt that the predominant pair of forces influencing the explosive growth in the number of transistors per chip has been the technological push of a continuous reduction in the cost per transistor or electronic function performed by a microchip coupled with the pull of ever-expanding markets and revenues.

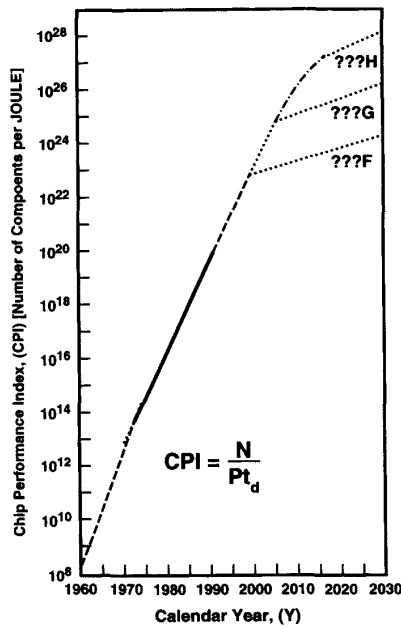


Fig. 31. Chip performance index  $CPI = N/Pt_d$  versus calendar year  $Y$ .

The paramount issue confronting these positive trends has been, is, and will be the concomitant exponential growth in the capital cost of a new high volume manufacturing line needed for each successive generation of microchips [63]. While this economic issue is well beyond the scope of the current discussion, one relevant hypothesis is explored.

The hierarchy of theoretical limits on microelectronics established over the past decade and more and summarized in this discussion does not indicate that the pessimistic or the realistic or even the optimistic projections of minimum feature size  $F$ , die area  $D^2$ , packing efficiency  $PE$ , number of transistors per chip  $N$ , and chip performance index  $N/Pt_d$  cannot be achieved. In other words, physical limits per se do not appear to be "show stoppers" over the next two decades. Moreover, assuming that the cost per electronic function performed by a microchip continues to decline, it does appear that market demand will continue to escalate over the next two decades simply because the capacity of the microchip to provide cost effective solutions to the myriad problems of the information revolution is virtually unlimited within this timeframe. Consequently, the paramount issue is unchanged: Will there continue to be sufficient economic incentives to risk the ever growing capital investments required for further reduction of the cost per function of microchips? It is feasible that the response will also be unchanged, especially if the manufacturing cost goals of Sematech are fulfilled [64]. Within the time interval addressed in this discussion, fundamental, material, device, circuit, and system physical limitations may well permit and virtually unbounded market opportunities may well stimulate development of the highly expensive manu-

facturing technology that will enable continuous reduction, although perhaps at a smaller than historic rate, in the cost per function of microchips. Consequently, it is imperative that we continue to pursue as deep an understanding as possible of the hierarchy of physical limits that govern future opportunities for GSI. The National Technology Roadmap for Semiconductors prepared under the leadership of Sematech, the Semiconductor Research Corporation, and the Semiconductor Industries Association is a laudable contribution toward this effort.

#### ACKNOWLEDGMENT

The author gratefully acknowledges the encouragement and support of Dr. Robert Burger and Dr. William Lynch in connection with SRC Contract 93-SJ-374. In addition, stimulating discussions with Dr. Vivek De over a period of years are sincerely appreciated. Finally, the author wishes to thank two anonymous reviewers recruited by the guest editor of this issue, Dr. Lewis Terman, for their constructive critiques of the original manuscript.

#### REFERENCES

- [1] J. D. Meindl, "The evolution of solid state circuits: 1958-1992-20??," 1993 *IEEE ISSCC Commemorative Suppl.*, pp. 23-26, Feb. 1993.
- [2] G. E. Moore, "Progress in digital integrated electronics," *IEEE IEDM Tech. Dig.*, pp. 11-13, 1975.
- [3] J. D. Meindl, "Theoretical, practical and analogical limits in ULSI," *IEEE IEDM Tech. Dig.*, pp. 8-13, 1983.
- [4] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. and Develop.*, vol. 5, no. 3, pp. 183-191, July 1961.
- [5] —, "Dissipation and noise immunity in computation and communication," *Nature*, vol. 335, pp. 779-784, Oct. 27, 1988.
- [6] E. Keonjian, Ed., *Micropower Electronics*. London and New York: Pergamon, 1964.
- [7] J. D. Meindl, *Micropower Circuits*. New York: Wiley, 1969.
- [8] M. Degrauwe et al., "Low power/low voltage: Future needs and envisioned solutions," 1994 *IEEE ISSCC Dig. Papers*, pp. 98-99.
- [9] S. Kohyama, "Semiconductor technology crises and challenges toward the year 2000," 1994 *Symp. VLSI Tech. Dig. of Papers*, pp. 5-8.
- [10] D. Singh, "Prospects for low power microprocessor design," *Proc. 1994 Int. Workshop on Low Power Design*, Napa, CA, Apr. 1994, p. 1.
- [11] S. Molhi and P. Chatterjee, "I-V microsystems-scaling on schedule for personal communications," *IEEE Circ. and Devices*, Mar. 1994, pp. 13-17.
- [12] A. P. Chandrakasan, S. Sheng, and R. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circ.*, vol. 27, pp. 473-484, Apr. 1992.
- [13] G. Moore et al., "Metal-oxide-semiconductor field effect devices for micropower logic circuitry," in *Micropower Electronics*, E. Keonjian, Ed. London/New York: Pergamon, 1964.
- [14] R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circ.*, vol. SC-7, pp. 146-152, Apr. 1972.
- [15] E. A. Vittoz, "Low-power design: Ways to approach the limits," 1994 *IEEE ISSCC Dig. Papers*, pp. 14-18.
- [16] J. D. Meindl et al., "Implantable telemetry," in *Methods of Animal Experimentation*, vol. 7, W. I. Gay and J. E. Heavner, Eds. New York: Academic, 1986, pp. 37-112.
- [17] P. H. Hudson and J. D. Meindl, "A monolithic micropower command receiver," *IEEE J. Solid-State Circ.*, vol. SC-7, pp. 125-134, Apr. 1972.
- [18] A. Chandrakasan, A. Burstein, and R. Brodersen, "A low power chipset for portable multimedia applications," 1994 *IEEE ISSCC Dig. Papers*, pp. 82-83.

- [19] R. W. Keyes, "Physical limits in digital electronics," *Proc. IEEE*, vol. 63, pp. 740–766, May 1975.
- [20] F. W. Sears, *Thermodynamics*. Reading, MA: Addison-Wesley, 1953.
- [21] H. Haken and H. C. Wolf, *Atomic and Quantum Physics*. Berlin: Springer-Verlag, 1984, pp. 83–85.
- [22] A. Bhavnagarwala, private communication.
- [23] M. V. Fischetti and S. E. Laux, "Monte carlo simulation of transport in technologically significant semiconductors-Part II: Submicrometer MOSFET's," *IEEE Trans. Electron Devices*, vol. 38, pp. 650–660, Mar. 1991.
- [24] F. Assaderaghi, "Observation of velocity overshoot in silicon inversion layers," *IEEE Electron Device Lett.*, vol. 14, pp. 484–486, Oct. 1993.
- [25] G. Baccarani *et al.*, "Generalized scaling theory and its application to a 0.25 micron MOSFET," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 452–470, Apr. 1984.
- [26] K. N. Ratnakumar and J. Meindl, "Short channel MOSFET threshold voltage model," *IEEE J. Solid-State Circ.*, vol. SC-17, pp. 937–947, Oct. 1982.
- [27] L. D. Yau, "A simple theory to predict the threshold voltage of short channel IGFET's," *Solid-State Electron.*, vol. 17, pp. 1059–1063, 1974.
- [28] R. R. Troutman, "VLSI limitations from drain induced barrier lowering," *IEEE Trans. Electron Devices*, vol. ED-26, pp. 461–469, 1979.
- [29] B. Agrawal, V. K. De, and J. D. Meindl, "Opportunities for scaling MOSFET's for GSI," *Proc. ESSDERC 1993*, pp. 919–926.
- [30] C. Fiegna *et al.*, "A new scaling method for 0.1–0.25 micron MOSFET," *May 1993 Symp. VLSI Tech. Dig.*, pp. 33–34.
- [31] C. Hu, "MOSFET scaling in the next decade and beyond," *Semicon Int.*, June 1994, pp. 105–114.
- [32] M. Ono *et al.*, "Sub-59nm gate length N-MOSFET's with 10nm phosphorous S/D junctions," *1993 IEEE IEDM Tech. Dig.*, pp. 119–121.
- [33] D. A. Antoniadis and J. E. Chung, "Physics and technology of ultra short channel MOSFET's," *1991 IEEE IEDM Tech. Dig.*, pp. 21–24.
- [34] D. Hisamoto *et al.*, "A fully depleted lean channel transistor (DELTA)-a novel vertical ultrathin SOI MOSFET," *IEEE Electron Device Lett.*, vol. 11, pp. 36–38, Jan. 1990.
- [35] D. J. Frank *et al.*, "Monte Carlo simulation of a 30nm dual gate mosfet: How short can Si go?" *1992 IEEE IEDM Dig. Papers*, pp. 553–556.
- [36] T. Tanaka *et al.*, "Ultrafast low power operation of P+N+ double-gate SOI MOSFETs," *1994 Symp. VLSI Tech. Dig. of Papers*, pp. 11–12.
- [37] K. Y. Toh *et al.*, "An engineering model for short-channel MOS device," *IEEE J. Solid-State Circ.*, pp. 950–958, Aug. 1988.
- [38] B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990, pp. 198–200.
- [39] J. Davis, private communication.
- [40] M. Nagata, "Limitations, innovations and challenges of circuits and devices into a half micrometer and beyond," *IEEE J. Solid-State Circ.*, vol. 27, pp. 465–472, Apr. 1992.
- [41] Y. Nakogame *et al.*, "An experimental 1.5v 64mb DRAM," *IEEE J. Solid-State Circ.*, vol. 26, pp. 465–472, Apr. 1991.
- [42] K. Ishibaski *et al.*, "A 1-V TFT-load SRAM using a two-step work-voltage method," *IEEE J. Solid-State Circ.*, vol. 27, Nov. 1992.
- [43] T. Sakata *et al.*, "Subthreshold-current reduction circuits for multi-gigabit DRAM's," *IEEE J. Solid-State Circ.*, vol. 29, pp. 761–769, July 1994.
- [44] K. Shimohigaski, "Low-voltage ULSI design," *IEEE J. Solid-State Circ.*, vol. 28, pp. 408–413, Apr. 1993.
- [45] Y. Nakagome, "Sub-1-V swing internal bus architecture for future low-power ULSI's," *IEEE J. Solid-State Circ.*, vol. 28, pp. 414–419, Apr. 1993.
- [46] Y. Taur *et al.*, "High performance 0.1 mm, CMOS devices with 1.5 V power supply," *IEEE IEDM Tech. Dig.*, pp. 127–130, Dec. 1993.
- [47] Y. Mu *et al.*, "An ultra low-power 0.1 um CMOS," *Symp. VLSI Tech. Dig.*, pp. 9–10, June 1994.
- [48] J. Burr and J. Shott, "A 200 m V self-testing encoder/decoder using standard ultra-low power CMOS," *IEEE ISSCC Dig.*, pp. 84–85, Feb. 1994.
- [49] H. B. Bakoglu and J. D. Meindl, "Optimal interconnection circuits for VLSI," *IEEE Trans. Electron Devices*, vol. ED-37, pp. 903–909, May 1985.
- [50] H. T. Kung, "Why systolic architectures," *IEEE Comput.*, pp. 37–46, Jan. 1982.
- [51] J. A. B. Fortes and B. W. Wah, "Systolic arrays—from concept to implementation," *IEEE Comput.*, pp. 12–17, July 1987.
- [52] H. S. Stone and J. Cocke, "Computer architecture in the 1990's," *IEEE Comput.*, pp. 30–38, Sept. 1991.
- [53] K. Chin *et al.*, "IBM enterprise system/9000 clock system: A technology and system perspective," *IBM J. Res. Develop.*, vol. 37, no. 5, pp. 867–874, Sept. 1992.
- [54] B. S. Landrum and R. L. Russo, "On a pin versus block relationship for partitioning of logic graphs," *IEEE Trans. Comput.*, vol. C-20, pp. 1469–1479, Dec. 1971.
- [55] W. E. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Trans. Circ. and Syst.*, vol. CAS-26, pp. 272–277, Apr. 1979.
- [56] A. Masaki, "Possibilities of deep-submicrometer CMOS for very high speed computer logic," *Proc. IEEE*, vol. 81, pp. 1311–1324, Sept. 1993.
- [57] G. A. Sai-Halasz, "High end processor trends and limits," *Proc. Interconnect Conf. on Advanced Microelectron. Device Proc.*, Sendai, Japan, pp. 753–760, Mar. 3–5, 1994.
- [58] J. C. Eble, private communication.
- [59] D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 4th ed. New York: Wiley, 1993.
- [60] S. G. Younis, "Asymptotically zero energy computing using split-level charge recovery logic," Ph.D. dissertation, Dept. of EECS, MIT, 1994.
- [61] *Proc. 1994 Int. Workshop on Low Power Design*, Napa, CA, Apr. 1994.
- [62] V. De, private communication.
- [63] C. R. Barrett, "Silicon valley, what next," *MRS Bull.*, pp. 3–10, July 1993.
- [64] W. J. Spencer, "National interests in a global semiconductor industry," *Distinguished Lecturer Series*, Georgia Inst. of Technol., Atlanta, GA, Oct. 3, 1994.
- [65] Semiconductor Ind. Assoc., *The National Technology Roadmap for Semiconductors*, 1994.



**James D. Meindl** (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Carnegie-Mellon University in 1955, 1956, and 1958, respectively.

He is presently the Joseph M. Pettit Chair Professor of Microelectronics at the Georgia Institute of Technology, Atlanta, GA. From 1986 to 1993 he was Senior Vice President for Academic Affairs and Provost of Rensselaer Polytechnic Institute, Troy, NY. From 1967 to 1986 he was with Stanford University, 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 Laboratory, and Founding Director of the Integrated Circuits Laboratory. He is a cofounder of Telesensory Systems. During 1965–1967, he was Director of the Integrated Electronics Division at the Fort Monmouth, NJ, US Army Electronics Lab. He authored a book on micropower circuits and over 300 technical papers on ultra-large scale integration, integrated electronics, and medical electronics. He edited *Brief Lessons in High Technology*.

Dr. Meindl is a fellow of the American Association for the Advancement of Science and a member of the American Academy of Arts and Sciences and the National Academy of Engineering and its Academic Advisory Board. He received the 1991 Benjamin Garver Lamme Medal from ASEE, the 1990 IEEE Education Medal, and the 1989 IEEE Solid-State Circuits Medal. At the 1988 IEEE International Solid-State Circuits Conference, he received the Beatrice K. Winner Award. In 1980 he received the IEEE Electron Devices Society's J. J. Ebers Award for contributions to the field of medical electronics. From 1970 to 1978, Dr. Meindl and his students received five outstanding paper awards at the IEEE International Solid-State Circuits Conference.