pathway DBs can impose an organizing framework on complex gene expression (or proteomics) data sets that facilitates their interpretation.

Future challenges for pathway DBs include modeling of large signaling networks in eukaryotic organisms; performing automated layout similar to that shown in Fig. 1 of the much larger pathway networks that exist in eukaryotic organisms, and supporting methods for user navigation through such a larger pathway network; defining standard ontologies for exchange of pathway data among different DBs and application programs; and creating new analysis algorithms for extracting new insights from pathway networks, such as to aid drug design by analyzing diseased human pathway networks, or predicting optimal drug targets for antimicrobial drug design.

One lesson for computer scientists provided by pathway DBs (and by other bioinformatics applications) concerns the importance of DB content to solving computational problems. Most computer scientists focus their attention on algorithms, thinking that the best way to solve a hard computational problem is through a better algorithm. However, for problems such as predicting the pathway complement of an organism from its genome, or predicting metabolic products that an organism can produce from a given growth medium, I know of no algorithms that can solve these problems without being coupled with an accurate and well-designed pathway DB.

By encoding scientific theories in a symbolic DB, scientists can more easily check those theories for internal consistency and for consistency with external data, can more easily refine theories that are found to violate external data, and can more easily assess the global properties of the system that such a theory describes. The genome revolution is increasing the need for pathway DBs in the biological sciences, and similar developments will occur in other sciences. However, effective implementation of this paradigm is hampered because most biologists (and most other scientists) receive essentially no education in DBs or knowledge representation. Although many scientists learn a computer programming language as part of their undergraduate education, introductory programming courses completely omit DB and knowledge representation concepts such as data models, ontologies, DB query languages, logical inference, DB design, and formal grammars—which explains why many biological DBs do not have a regular syntactic structure, much less a consistent or precisely defined semantics. As science enters the information age, it is crucial that the computer-science education that scientists receive covers symbolic computing as well as numerical computing.

References and Notes

1. C. Ouzounis, P. Karp, *Genome Res.* **10**, 568 (2000).
2. P. Karp *et al.*, *Nucleic Acids Res.* **28**, 56 (2000).
3. P. Karp, in *Nucleic Acid and Protein Databases and How To Use Them* (Academic Press, London, 1999), pp. 269–280..
4. F. Blattner *et al.*, *Science* **277**, 1453 (1997).
5. M. Kanehisa, S. Goto, *Nucleic Acids Res.* **28**, 27 (2000).
6. R. Overbeek *et al.*, *Nucleic Acids Res.* **28**, 123 (2000).
7. L. Ellis, C. Hershberger, L. Wackett, *Nucleic Acids Res.* **28**, 377 (2000).
8. P. D. Karp, *Trends Biochem. Sci.* **23**, 114 (1998).
9. P. Karp, M. Krummenacker, S. Paley, J. Wagg, *Trends Biotechnol.* **17**, 275 (1999).
10. P. D. Karp, *Bioinformatics* **16**, 269 (2000).
11. H. Jeong, S. P. Mason, A.-L. Barabasi, Z. N. Oltvai, *Nature* **411**, 41 (2001).
12. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabasi, *Nature* **407**, 651 (2000).
13. P. Karp, C. Ouzounis, S. Paley, in *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, D. States, P. Agarwal, T. Gaasterland, L. Hunter, R. Smith, Eds. (American Association for Artificial Intelligence, Menlo Park, CA, 1996).
14. C. Schilling, B. Palsson, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4193 (1988).
15. J. Edwards, B. Palsson, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5528 (2000).
16. P. Romero, P. Karp, in *Proceedings of the Pacific Symposium on Biocomputing*, R. Altman, T. Klein, Eds. (World Scientific, Singapore, 2001), pp. 471–482.
17. T. Winograd, in *Representation and Understanding* (Academic Press, New York, 1975), pp. 185–210.

18. J. Collado-Vides, I. Paulsen, M. Riley, and M. Saier are collaborators on the EcoCyc project. J. Collado-Vides assisted in the analysis of the *E. coli* genetic network. S. Paley produced Fig. 2. C. Ouzounis, T. Garvey, A. Rzhetsky, and I. Sim provided valuable comments on this manuscript. The development of EcoCyc, MetaCyc, and the Pathway Tools has been funded by grant 1-R01-RR07861-01 from the Comparative Medicine Program of the NIH National Center for Research Resources.

VIEWPOINT

Limits on Silicon Nanoelectronics for Terascale Integration

James D. Meindl,* Qiang Chen, Jeffrey A. Davis

Throughout the past four decades, silicon semiconductor technology has advanced at exponential rates in both performance and productivity. Concerns have been raised, however, that the limits of silicon technology may soon be reached. Analysis of fundamental, material, device, circuit, and system limits reveals that silicon technology has an enormous remaining potential to achieve terascale integration (TSI) of more than 1 trillion transistors per chip. Such massive-scale integration is feasible assuming the development and economical mass production of double-gate metal-oxide-semiconductor field effect transistors with gate oxide thickness of about 1 nanometer, silicon channel thickness of about 3 nanometers, and channel length of about 10 nanometers. The development of interconnecting wires for these transistors presents a major challenge to the achievement of nanoelectronics for TSI.

Silicon technology has advanced at exponential rates in both performance and productivity throughout the past four decades. From 1960 to 2000, the energy transfer associated with a binary switching transition—the canonical digital computing operation—decreased by about five orders of magnitude and the number of transistors per chip increased by about nine orders of magnitude. Such exponential advances must eventually come to a halt imposed by a hierarchy of physical limits. The five levels of this hierarchy are defined as fundamental, material, device, circuit, and system (*1*). A coherent analysis of the key limits at each of these levels reveals that silicon technology has an enormous remaining potential to achieve TSI of more than 1 trillion transistors per chip, with critical device dimensions or channel lengths in the 10-nm range. This potential represents more than a three-decade increase in the number of transistors per chip and more than a one-decade reduction in minimum transistor feature size compared with the state of the art in 2001. Fundamental physical limits that are independent of the characteristics of any particular material, device structure, circuit configuration, or system architecture are virtually impenetrable barriers to future advances of TSI.

Binary switching transitions implemented with transistors are indispensable to performing computation in a digital system. The energy transfer per binary transition is a revealing metric for comparing the performance of

School of Electrical and Computer Engineering, Microelectronics Research Center, Georgia Institute of Technology, Atlanta, GA 30332–0269, USA.

*To whom correspondence should be addressed. E-mail: james.meindl@mirc.gatech.edu

switching operations at all levels of the hierarchy. Consider the power-delay plane, where the ordinate is the average power transfer, P, during a binary transition and the abscissa is the time interval of the transition, t_d. The use of logarithmic scales on both axes results in a diagonal line (or locus) where the switching energy, $E = Pt_d$, remains constant. During the past four decades, constant switching energy loci have migrated continuously toward the lower left corner of the power-delay plane, reflecting a monotonically decreasing binary switching energy (*1*). The prime cause of this migration has been the scaling down of the dimensions of transistors and their binary signal voltage swing, typically equal to the supply voltage. Supply voltage is reduced to maintain a nearly constant electric field (in V/cm) or electrical stress on the transistor. Scaling of transistors reduces their energy dissipation per binary transition, their intrinsic switching delay, their area, and therefore their cost.

The second indispensable function performed in a digital system is communication, implemented by interconnects or wires. The primary purpose of an interconnect is communication between distant points with small latency. Interconnect performance can be elucidated at all levels of the hierarchy by plotting the square of the reciprocal interconnect length, L^{-2}, against latency, τ. In the L^{-2} versus τ plane, with logarithmic scales on both axes, a diagonal line is a locus of constant value of $L^{-2}\tau = r_{int}c_{int}$ s/cm^2 or constant distributed resistance-capacitance product. This product is the prime figure of merit for interconnects. During the past four decades, constant distributed resistance-capacitance loci have migrated continuously toward the upper right corner of the L^{-2}-τ plane, reflecting a continuously increasing distributed resistance-capacitance product and consequently a larger latency for communication between two fixed points. Larger latency cannot be avoided because the cross-sectional dimensions of interconnects must be scaled down to provide the dense wiring required by smaller and smaller transistors. Consequently, during the past decade, interconnect latency (as well as energy dissipation) has become a primary constraint on current gigascale integration. Exploring key limits at each of the five levels of the hierarchy in the power-delay and reciprocal length squared-latency planes elucidates future opportunities for TSI.

Fundamental Limits

The three key fundamental limits on TSI are derived from thermodynamics, quantum mechanics, and electromagnetics (*1*, *2*). The fundamental limit on signal energy transfer during a binary switching transition is $E(\min) = (\ln 2)kT$, where k is Boltzmann's constant and T is absolute temperature. This limit is characterized as fundamental because its value is independent of the properties of any particular material, device, or circuit that may be used to implement the binary transition (*3*). Its importance as a constraint on nanoelectronics for TSI is unsurpassed. In simple physical terms, the limit reveals that a single electron undergoing a binary transition must have an energy comparable to its thermal energy, $(3/2)kT$, to satisfy the quintessential requirement of binary signal discrimination.

The first statement of this limit known to the authors is attributed to John von Neumann, who "computed the thermodynamical minimum of energy per elementary act of information from the formula $kT\log_e N$" where $N = 2$ for a binary act (*4*, p. 183). Keyes observes, however, that "the report of von Neumann's ideas fails to provide any justification of this assertion or explanation of the reasoning underlying it" (*5*). Landauer derived the same result by analyzing a hypothetical binary device consisting of a particle in a bistable potential well (*5*). On the basis of earlier work of Swanson and Meindl (*6*), the minimum switching energy of an ideal transistor operating in the simplest digital circuit, an inverter, is $E(\min) = (\ln 2)kT$ (*3*). Precisely the same result is derived (*3*) by treating an isolated interconnect as a communication channel described by Shannon's classical theorem for channel capacity (*7*). This fundamental limit receives further support from the observation that on the basis of a Boltzmann probability density function, the probability of error is 0.5 for a binary transition with signal energy transfer $E(\min) = (\ln 2)kT$ (*8*).

Quantum mechanics and, more specifically, the Heisenberg uncertainty principle (*9*) define the second fundamental limit, which requires a signal switching energy transfer $\Delta E \geq h/t_d$, where h is Planck's constant and t_d is the transition time. This limit results from the wave nature of the electron and the resulting uncertainty in its position-momentum and energy-time relations (*9*).

The fundamental limits based on thermodynamics and quantum mechanics result in a "forbidden region" in the power-delay plane (red region, Fig. 1). In this region, no binary transition can operate, regardless of the materials, devices, or circuits used for its implementation.

The third fundamental limit from electromagnetics simply expresses the fact that the time of flight, τ, of an electromagnetic wave traveling along any metallic interconnect or optical fiber of length L is strictly limited by the velocity of light in free space, c_0, according to $\tau \geq L/c_0$ (Fig. 2) (*1*). The red region is again a forbidden zone of operation for any interconnect regardless of the materials or structure used for its implementation.

Material Limits

Material limits are determined by the properties of the particular semiconductor, dielectric, and metallic materials used but must be essentially independent of the structural features and dimensions of particular devices (*1*, *10*). There are five key material limits. Silicon imposes four of them: a switching energy, a transit time, a thermal conductance, and a dopant fluctuation limit. The dielectric constant of the insulator of a multilevel interconnect network imposes the final material limit.

The switching energy limit is determined by the amount of energy E that must be stored in a cube of semiconductor material to support a selected binary transition voltage, V_o. This is the voltage applied between two opposite faces of the cube in the direction of current flow. The expression for this limiting energy is given by $E = \varepsilon(V_o)^3/2\mathcal{E}_c$, where ε is the permittivity and $\mathcal{E}_c$ is the breakdown electric field strength of the semiconductor material. The transit time limit t_d is defined by the smallest time interval required for an electron to be transported through the cube. This limiting time is expressed as $t_d = V_o/v_s\mathcal{E}_c$, where v_s is the electron saturation velocity (the largest possible electron velocity whose value is 10^7 cm/s in silicon) in a particular material. The thermal conductance limit defines the maximum amount of power, P, that may be dissipated in a single transistor within a particular semiconductor chip. P must equal the rate of heat removal under steady state conditions. The power dissipation limit is given by $P = \pi K v_s \Delta T t_d$, where K is the thermal conductivity, v_s is the saturation velocity of the semiconductor material, ΔT is the temperature difference between the transistor and an ideal heat sink for the semiconductor chip, and t_d is the device transit time.

The minimum binary transition voltage V_o needed for high-performance devices and circuits for TSI is believed to be 0.5 V. The orange region defined by the switching energy, transit time, and thermal conductance limits (Fig. 1) is a second forbidden zone of operation, imposed by the material limits of silicon. No silicon transistor regardless of its structural features can operate in this orange forbidden region. It is especially notable that the three expressions defining the material limits are essentially independent of the structural features and dimensions of any particular device. A rare exception may be certain very small devices exhibiting an effective increase in carrier velocity due to a short-range phenomenon termed velocity overshoot (*11*).

The fourth key semiconductor material limit is a dopant fluctuation limit, which is defined by the expression $\sigma/\mu = (\ell/\Delta\chi)^{3/2}$. The standard deviation and the mean value of the number of dopant atoms within a cube of semiconductor material of dimension $\Delta\chi$ are

σ and μ, respectively; ℓ is the average distance between dopant atoms in the cube. This expression reveals that the standard deviation of the number of dopant atoms in a cube of semiconductor material, σ, increases without bound as the cube dimension, $\Delta\chi$, decreases. This poses a critical concern for TSI because it hints that deviations in the values of key device parameters, such as the threshold voltage of a transistor, may increase without bound as device dimensions are scaled to the 10-nm range.

The time of flight τ of an electromagnetic wave in a solid dielectric material with a relative permittivity, ε_r, is expressed by $\tau = L/(\varepsilon_r)^{1/2}c_0$, which defines the fifth key material limit. The dashed locus (Fig. 2) represents this limit for $\varepsilon_r = 2$. The orange zone is a forbidden region for any interconnect whose relative permittivity ε_r is greater than 2. Relative permittivity values less than two generally require porous materials consisting of gas "balloons" encased by thin solid walls.

Device Limits

There are five key limits at the device level (*1*, *12*) of the hierarchy. Metal-oxide-semiconductor field effect transistors (MOSFETs), the most critical devices of TSI, impose a switching energy, a transit time, and a parameter fluctuation limit. Interconnects impose key latency and cross-talk limits. An advanced MOSFET structure is illustrated in Fig. 3.

During a binary switching transition, the energy stored on the capacitive gate or control electrode of a MOSFET device is transferred. This energy therefore represents its switching energy limit, given by $E = (1/2)C_g(V_{dd})^2$. The gate capacitance of a minimum geometry MOSFET is expressed by $C_g = \varepsilon_{ox}(L_{ch})^2/T_{ox}$, where ε_{ox} is the permittivity of the gate oxide, L_{ch} is the channel length, and T_{ox} is the gate oxide thickness. The binary signal voltage swing is assumed to equal the supply voltage V_{dd}, as is the case for the predominant complementary metal-oxide-semiconductor (CMOS) digital circuit family. The lower limit on E corresponds to a minimum channel length, L_{ch}, or minimum size MOSFET operating at a minimum supply voltage, V_{dd}.

The intrinsic switching delay of a MOSFET can be expressed in its simplest form as the transit time of carriers across its channel from source to drain or $t_d = L_{ch}/v_s$, where the average velocity of a transiting electron is taken to be the saturation velocity, v_s.

Both the switching energy, E, and the switching delay, t_d, of a MOSFET will be at a minimum for the smallest possible channel length, L_{ch}. It is this observation that has driven the quest for ever smaller transistors for the past four decades. Unfortunately, as transistor channel length is scaled down, eventually the gate or threshold voltage at which the device switches from open or nonconducting to closed or strongly conducting precipitously decreases. The double-gate MOSFET structure (Fig. 3) enables the smallest values of channel length. In this device, drain-to-source channel current is controlled by electric fields created by both top and bottom gate voltages rather than from a top gate only as in conventional MOSFETs (*1*).

A recently derived solution to the two-dimensional Poisson equation of electrophysics defines the channel length of a double-gate MOSFET as (*13*)

$$L_{ch} = 2\lambda \ln\left[\frac{4\lambda}{T_{Si}}\frac{\left(\sin\frac{T_{Si}}{4\lambda} + \frac{T_{Si}}{r\lambda}\cos\frac{T_{Si}}{4\lambda}\right)}{\frac{1}{r} + \frac{1}{2} + \frac{1}{2}\left(\frac{T_{Si}}{r\lambda}\right)^2}\left(1 - \frac{\ln 10}{\beta S}\right)^{-1}\right] \quad (1)$$

where $\lambda = [1 + (1/r)][1 + (\pi/2)]^{-1}T_{Si}$ and $r = T_{Si}/3T_{ox}$, and $\beta = q/kT$. Figure 4 illustrates two plots of Eq. 1, which indicate the key opportu-

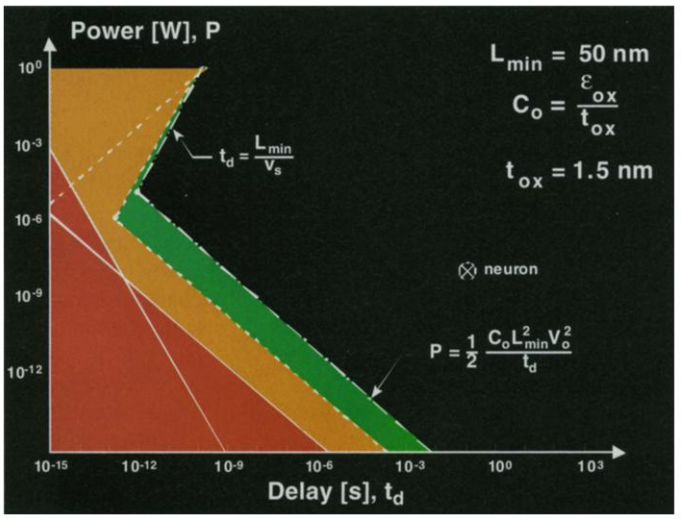

Fig. 1. Average power transfer during a binary transition, *P*, versus transition time, t_d, for the first three levels of the hierarchy. The red, orange, and green zones are forbidden by fundamental, silicon material, and 50-nm channel length transistor device level limits, respectively.

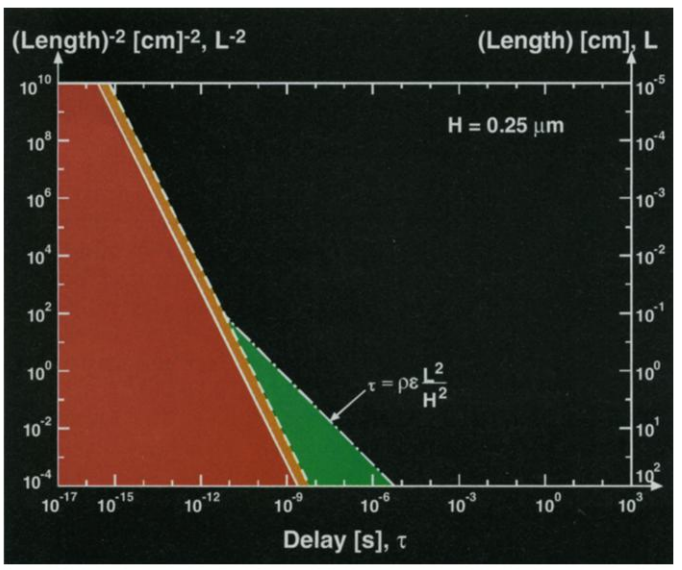

Fig. 2. Reciprocal interconnect length squared, L^{-2}, versus latency, τ, for the first three levels of the hierarchy. The red, orange, and green zones are forbidden by fundamental, material (ε_r = 2.0), and 250-nm-wide interconnect device level limits, respectively.

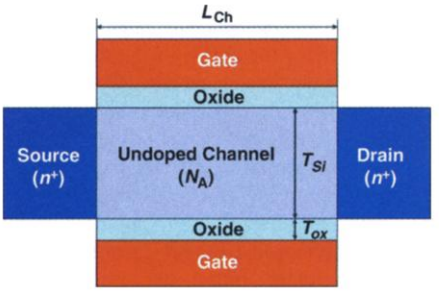

Fig. 3. Schematic diagram of the cross section of a symmetrical double-gate MOSFET. The gate electrode is highly conducting, the gate oxide is highly insulating, and the undoped channel is semiconducting silicon. In this so-called metal-oxide-semiconductor field effect transistor, or MOSFET, an input signal voltage applied between the gate and source electrodes controls output current flow from drain to source.

nity for double-gate MOSFET channel lengths in the 10-nm range. The ultimate challenge of TSI is implementing several trillion of these devices—with tightly controlled gate oxide thickness T_{ox} in the 1.0-nm range, silicon channel thickness T_{Si} in the 3.0-nm range, and channel length L_{ch} in the 10-nm range—in a single silicon chip selling for less than $100. As indicated in Fig. 4, these values of T_{ox} and T_{Si} are necessary to achieve channel lengths L_{ch} in the 10-nm range.

The third key device limit concerns the need for ultratight control of MOSFET dimensions and dopant impurity concentrations to preclude parameter fluctuations so large as to cause functional faults in device and circuit operation. Random deviations from nominal values of MOSFET and interconnect parameters preclude attainment of the precise performance levels defined by the hierarchy of limits on TSI. A prime example of this generalization is the fundamental limit imposed by thermodynamics on signal energy transfer during a binary switching transition, $E(\min) = (\ln 2)kT$. At this level of signal energy transfer, the probability of error during a binary transition is unacceptably high and therefore mandates a larger value of switching energy and its associated lower probability of error. Moreover, double-gate MOSFET models of the impact of random placement of dopant atoms in the channel region (Fig. 3) reveal that control of threshold voltage deviation demands the use of very lightly doped (typically $< 10^{15}$ atoms/cm^3) channel regions (*14*).

A distributed resistance-capacitance network serves as the model for an isolated interconnect whose response time or latency, τ, increases quadratically as interconnect length increases and as metal width and height as well as insulator thickness are scaled downward to increase wiring density (*1*). (As the width and height of a metal interconnect continue to scale downward, an additional severe deleterious effect enters the problem. This is the increase in the effective resistivity, ρ, that results from several factors, including strong electron scattering at the interface of the conductor and its surrounding insulator, and from large temperature increases resulting from the poor thermal conductivity of insulating layers.)

The normalized peak cross-talk voltage due to capacitive coupling between a quiescent interconnect and two adjacent parallel interconnects that undergo binary switching transitions is given by $V_n/V_{dd} = (1/2)[c_m/(c_{int} + c_m)]$, where c_m is the distributed mutual capacitance between the quiescent interconnect and an adjacent interconnect. As mutual capacitance increases because of smaller interconnect spacing, peak cross-talk voltage increases (*15*).

The MOSFET switching energy and transit time limits result in the green forbidden zone of operation for a conventional (that is, single-gate bulk silicon) device whose channel length is greater than 50 nm (Fig. 1). A 50-nm channel length represents a conservative value for limiting channel length of such MOSFETs. The latency of an interconnect modeled as a distributed resistance-capacitance network is illustrated in Fig. 2. The green region represents a forbidden zone of operation for any interconnect with a copper conductor, an insulator with a relative permittivity of two, and a square cross-sectional dimension of 250 nm (a suitable value for intermediate length interconnects). Figures 1 and 2 illustrate the comparative values of key limits at the first three levels of the hierarchy (*1*).

Circuit Limits

The six key circuit limits (*1*, *12*) on TSI are a static transfer curve, a switching energy, and a propagation delay limit imposed by CMOS logic circuits; latency and signal contamination limits imposed by global interconnect circuits; and a performance fluctuation limit.

To provide the quintessential capability of binary signal discrimination, the signal voltage swing of a CMOS digital logic circuit must satisfy the constraint $V_{dd} \geq 2(\ln 2)kT/q \geq 0.038$ V, where q is the charge of a single electron and $T = 300$°C (*8*). This static transfer curve limit applies to the predominant static CMOS logic circuit family for which binary signal swing is equal to the supply voltage. The switching energy limit is determined by the amount of energy that is transferred during a binary transition of an inverter, the basic circuit of the CMOS logic family. The switching energy is given by $E = (1/2)C_c(V_{dd})^2$, where C_c is the capacitance loading the output terminals of the circuit (*1*). The propagation delay limit is the average time, t_d, required for a binary signal appearing at the input terminals of a logic circuit to be propagated to its output terminals. In essence, t_d is simply the circuit latency (*1*).

The latency of a global interconnect circuit is, for example, the time required for a signal to propagate from the output terminals of a driver circuit, feeding a global interconnect extending from corner to corner of a chip, to the input terminals of a receiver circuit. This latency is minimal if the total resistance of the interconnect is small compared with its characteristic impedance, Z_o, and the output resistance of the driver equals Z_o (*1*). The characteristic impedance is given by $Z_o = (L/C)^{1/2}$, where L and C are the distributed inductance and capacitance per unit length of the interconnect, respectively.

The signal contamination limit results from mutual inductance and capacitance between a global interconnect, the victim, and its surrounding interconnects, the aggressors, causing unwanted or contaminating noise to appear on the victim when an intended signal appears on the aggressors. A simplified expression for the normalized peak cross-talk noise is given by $V_n/V_{dd} = (\pi/4)[c_m/(c_{int} + c_m)]$, where c_m is the mutual capacitance between adjacent interconnects and c_{int} is the capacitance between an interconnect and its underlying conducting plane (*15*, *16*).

The performance fluctuation limit at the circuit level results from transistor and interconnect electrical parameters deviating from their nominal values for whatever reasons including intrinsic and extrinsic manufacturing tolerances, temperature variations, supply voltage changes, and so forth. As previously

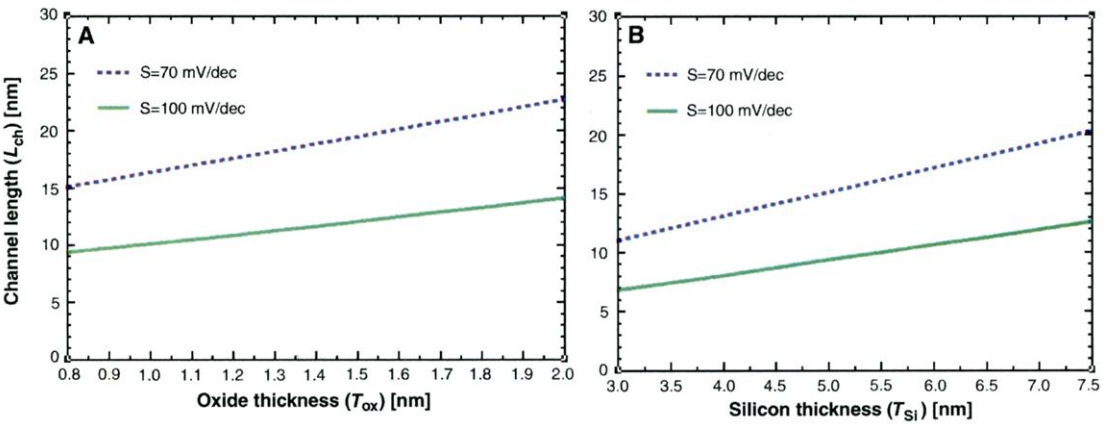

Fig. 4. (**A**) Channel length versus oxide thickness for $T_{Si} = 5$ nm. (**B**) Channel length versus silicon thickness for $T_{ox} = 0.8$ nm. These curves illustrate the potential to achieve double-gate MOSFETs with 10-nm channel lengths for gate oxide thickness in the 1.0-nm range and silicon channel thickness in the 3.0-nm range.

noted, fluctuations prevent circuit performance levels from reaching those defined by nominal physical limits. Typical increases in propagation delay and power dissipation due to such fluctuations are 30 and 50% above nominal for 50-nm generation CMOS logic circuits (*17*).

The switching energy and propagation delay limits for 50-nm generation CMOS logic circuits are illustrated in Fig. 5; Fig. 6 illustrates the global interconnect latency limit. In both figures, the blue regions define forbidden zones for operation due to circuit limits.

System Limits

Architecture, switching energy, heat removal, clock frequency or timing, and chip size impose five critical system limits on TSI. To elucidate these limits, it is helpful to select a representative set of requirements that must be satisfied by a gigascale system. The system to be considered requires 1 billion logic gates implemented with 50-nm generation CMOS technology. The required heat removal capacity of the package must not exceed 50 W/cm^2. The required clock frequency is 10 GHz. The entire system must be fabricated within a single silicon chip.

A distributed shared memory multiprocessor architecture that consists of a 24 by 24 array of 576 identical macrocellular microprocessors each containing 1.73 million gates is assumed. Each macrocell communicates directly only with its four nearest neighbors. The relatively small size of a macrocell and its nearest-neighbor-only external interconnects ensure relatively short internal and external interconnects and therefore small interconnect capacitances and hence small latency and switching energy dissipation.

To determine the switching energy limit, it is necessary to derive the complete stochastic interconnect length distribution of a macrocell (*18*). This enables calculation of the average capacitance, C_s, loading a two-input CMOS logic gate in the critical path of a macrocell. The switching energy limit is given by $E = (1/2)C_s(V_{dd})^2$, where V_{dd} is determined by minimizing the sum of the switching and static energy dissipation during a clock cycle (*19*).

The heat removal limit requires that the total power dissipation of the chip, P_t, is less than the cooling capacity of the package or $P_t \leq QA$, where Q is the cooling coefficient of the package (in W/cm^2) and A is the chip area. Heat removal actually limits the performance or maximum clock frequency of the chip (*1*).

The clock frequency limit requires that the clock period, T_c, must be greater than the sum of the clock skew, T_{cs}, and the critical path delay, T_{cp}, or $T_c \geq T_{cs} + T_{cp}$. Clock skew is the maximum difference in arrival times of a clock pulse at any two locations on the chip, and critical path delay is the maximum time interval required for a signal to propagate between two clocked locations.

The interior of the tiny white triangle in the P-t_d plane of Fig. 5 is the allowable design space for a system that fulfills all of its specified critical requirements. The surrounding purple region is a forbidden zone of operation in which one or more critical requirements cannot be fulfilled. Similarly, the small white triangle in the L^{-2}-τ plane of Fig. 6 represents the allowable design space and the purple zone is a forbidden region. The orthogonal sides of the triangle in the Fig. 6 are defined by the edge length of a macrocell and the latency of an interconnect of the same length.

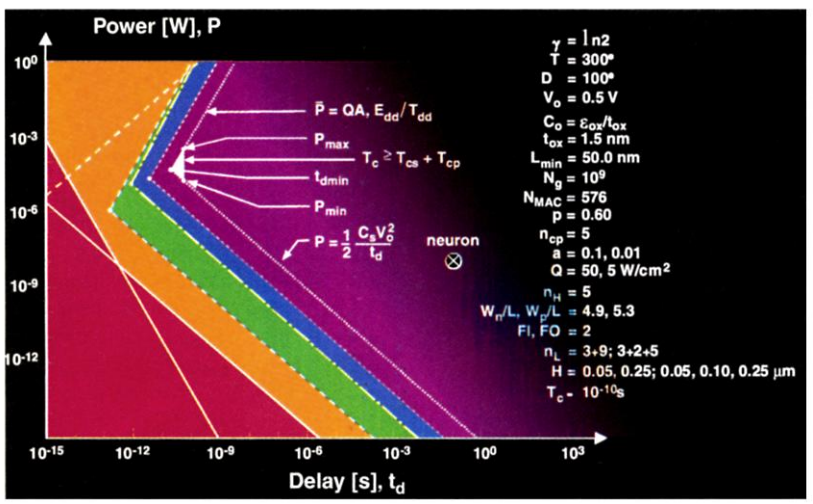

Fig. 5. P versus t_d for all levels of the hierarchy. The blue and purple zones are forbidden by representative gigascale circuit and system limits. The tiny white triangle is the allowable design space for a representative gigascale chip.

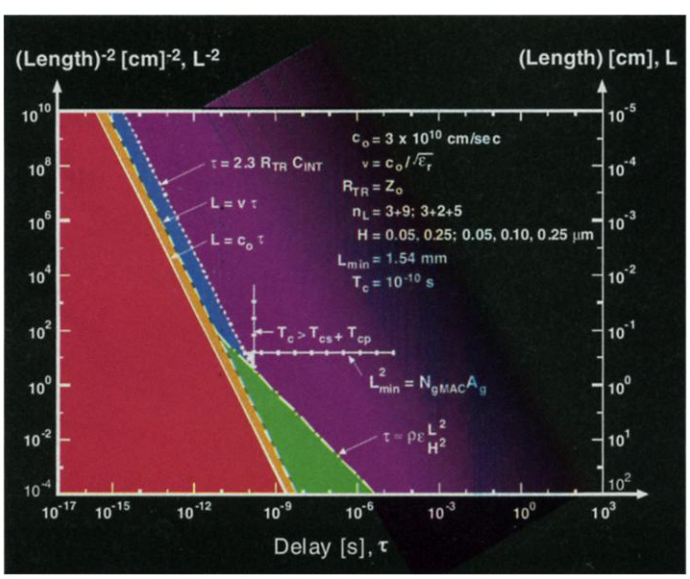

Fig. 6. L^{-2} versus τ for all levels of the hierarchy. The blue and purple zones are forbidden by representative gigascale interconnect circuit and system level limits. The tiny white triangle is the allowable design space for the longest interconnects of a representative gigascale chip.

Conclusion

A hierarchy of fundamental, material, device, circuit, and system limits reveals that 10-nm TSI is feasible assuming the critical development of double-gate MOSFETs with gate oxide thickness in the 1.0-nm range, silicon channel thickness in the 3.0-nm range, and channel length in the 10-nm range.

In Fig. 5, the white triangle—the allowable design space for a year 2011 generation TSI system (*20*)—is separated from the forbidden red zone imposed by fundamental limits by over five orders of magnitude. This is observed by noting the separation of the loci, representing the fundamental limit from thermodynamics and the system switching energy limit, along the abscissa of the figure. This huge separation is the result of the large interconnect capacitance that must be charged or discharged during a binary transition and the relatively large binary signal swing of 0.5 V. This amount of signal swing is necessary for large drive currents, leading to small circuit propagation delays and hence 10-GHz clock frequencies.

After four decades of rapid advances in both the performance and productivity of silicon semiconductor technology, a systematic assessment of its hierarchy of physical limits reveals an enormous remaining potential to

advance from current multibillion transistor chips to the multitrillion transistor range of terascale integration.

References

1. J. D. Meindl, *Proc. IEEE* **83** (no. 4), 619 (1995).
2. R. W. Keyes, *Proc. IEEE* **89** (no. 3), 227 (2001).
3. J. D. Meindl, J. A. Davis, *IEEE J. Solid State Circuits* **35** (no. 10), 1515 (2000).
4. J. von Neumann, *Theory of Self-Reproducing Automata* (Univ. of Illinois Press, Urbana, IL, 1966), p. 66.
5. R. W. Keyes, *Proc. IBM J. Res. Dev.* **5**, 183 (1961).
6. R. M. Swanson, J. D. Meindl, *IEEE J. Solid State Circuits* **SC-7**, 1146 (1972).
7. C. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).
8. J. D. Meindl, *Proc. IEEE* **89** (no. 3), 223 (2001).
9. H. Haken, H. C. Wolf, *Atomic and Quantum Physics* (Springer-Verlag, Berlin, 1984), chap. 7.
10. J. D. Plummer, *Proc. IEEE* **89** (no. 3), 240 (2001).
11. J. B. Roldan, F. Gamiz, J. A. Lopez-Vallanueva, J. E. Carceller, P. Cartujo, *Semicond. Sci. Technol.* **12**, 321 (1997).
12. D. J. Frank *et al.*, *Proc. IEEE* **89** (no. 3), 259 (2001).
13. Q. Chen, private communication.
14. X. Tang, V. K. De, L. Wang, J. D. Meindl, *Proc. IEEE Int. SOI Conf.* **1999**, 42 (1999).
15. J. Davis, J. D. Meindl, *IEEE Trans. Electron Devices* **47**, 2068 (2000).
16. ———, *IEEE Trans. Electron Devices* **47**, 2078 (2000).
17. K. A. Bowman, X. Tang, J. C. Eble, J. D. Meindl, *IEEE Trans. Electron Devices* **45** (no. 3), 580 (1998).
18. J. A. Davis, V. K. De, J. D. Meindl, *IEEE Trans. Electron Devices* **45** (no. 3), 590 (1998).
19. A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, J. D. Meindl, *IEEE Trans. Very Large Scale Integration (VLSI) Syst.* **8**, 235 (2000).
20. *The International Technology Roadmap for Semiconductors (ITRS)* (Semiconductor Industry Association, San Jose, CA, 1999).

VIEWPOINT

Blazing Pathways Through Genetic Mountains

David K. Gifford

It is now widely accepted that high-throughput data sources will shed essential understanding on the inner workings of cellular and organism function. One key challenge is to distill the results of such experiments into an interpretable computational form that will be the basis of a predictive model. A predictive model represents the gold standard in understanding a biological system and will permit us to investigate the underlying cause of diseases and help us to develop therapeutics. Here I explore how discoveries can be based on high-throughput data sources and discuss how independent discoveries can be assembled into a comprehensive picture of cellular function.

To date, most discoveries that have been based on expression data have relied on data visualization. For example, in this issue, Kim *et al.* describe the first large compendium of *Caenorhabditis elegans* expression data (*1*). The 533 microarray experiments discussed characterize the transcriptome of *C. elegans* cells in a wide variety of growth conditions, developmental stages, and genetic backgrounds. The coexpression of genes in these experiments gives important information about potential gene coregulation and the functions of previously uncharacterized genes in *C. elegans*. Thus, these data will be an important basis for further research in the *C. elegans* community.

Kim *et al.* visualize the *C. elegans* expression data in three dimensions for analysis. Groups of related genes in this three-dimensional approach appear as mountains, and the entire transcriptome appears as a mountain range. Distances in this synthetic geography are related to gene similarity, and mountain heights are related to the density of observed genes in a similar location. A three-dimensional approach is a departure from the common practice of analyzing expression data in a single dimension. Single-dimension analysis places genes in a total ordering, limiting our ability to see important relationships.

Department of Computer Science, Massachusetts Institute of Technology, 200 Technology Square, Cambridge, MA 02139, USA. E-mail: gifford@mit.edu

Visualization-based approaches are an important first step toward understanding cellular function. Expression visualization allows us to hypothesize potential gene-gene relationships that can be experimentally tested. For example, when a visualization tool shows that genes are coexpressed, it is natural to search for transcriptional activators that are shared between the genes. The results of such searches are typically expressed in schematic form, with the schematics depicting how genes influence one another's expression and activity. Often posttranslational modifications of proteins play a large role in their activities, and these modifications must also be captured in a schematic diagram to accurately predict the behavior of a system.

The individual elements of understanding that grow out of visualization and subsequent experiments can be naturally organized into a model-based approach to discovery. Model-based approaches codify our understanding of the underlying causes of data variation that is observed in data visualization, and the integration of results into a system model is necessary for broad understanding and insight. In a model-based approach, competing models that describe a function are constructed, and the models are scored against experimental data. The score of a model describes the likelihood of observing the experimental data given the model under consideration. Thus, models provide a principled way of judging the relative likelihood of competing hypotheses. When many models have roughly the same score, it is possible to determine the features that they share in common. The shared features of high-scoring models represent biological relationships that are likely to be important.

Despite the extraordinary discriminatory benefits of models, many biologists retreat from this approach with concerns about complex differential equations, unintelligible computer commands, and a feeling of unease that researchers will not be able to grasp the subtleties of what the models are saying. Furthermore, many model-based approaches require the values of reaction parameters that we do not yet know and that are difficult to approximate from contemporary high-throughput data sources. New approaches to modeling that are intuitive, can capture high-level structure, and are parameter-free would overcome these problems and motivate more biologists to capture and analyze in computational form what they suspect to be true.

Structured computational models, and in particular graphical models, have recently been proposed as a parameter-free approach for modeling biological network structure (*2*, *3*). Just like the schematic diagrams familiar to biologists, a graphical model captures the qualitative relationships between variables. Vertices in a graphical model represent variables such as mRNA expression levels, protein levels, environmental conditions, genotype, and phenotype. Edges in a graphical model describe relationships between variables and can be annotated with typical biological semantics, such as enhances or represses.

Once constructed, a graphical model represents both a conceptual understanding of a biological system and a computational means for predicting the effects of pertubations to the system. For example, Fig. 1 illustrates how a graphical model can explain data in a form that is simpler and more easily interpretable compared with conventional clustering di-