Scanning the Issue

Special Issue on Limits of Semiconductor Technology

Throughout the past four decades, semiconductor technology has been advancing at exponential rates in both productivity and performance. In the real world, such exponential advances do not continue endlessly. Therefore, the purpose of this special issue is to present a systematic assessment of early 21st century opportunities for gigascale and perhaps terascale integration. The central thesis of this assessment is that a hierarchy of limits whose five levels are codified as: 1) fundamental; 2) material; 3) device; 4) circuit; and 5) system [1] will govern future opportunities for gigascale integration (GSI). At each level of this hierarchy, two categories of limits are addressed: theoretical and practical. Theoretical limits are informed by the law's of physics, which change very rarely, and by technological invention, which occasionally causes very rapid change. Practical limits must be in compliance with physical constraints, but must also take into account manufacturing process, material, equipment, facilities, and labor costs as well as markets.

The fact that limits on GSI are subject to the laws of physics, technological invention, and economic factors is an unmistakable signal that the derived value of any particular limit is subject to the initial assumptions of its derivation. Moreover, the diversity and number of initial assumptions that must be engaged escalates with the level of the hierarchy. Definition of the value of a *fundamental* limit may require only the specification of operating temperature, whereas derivation of a meaningful monolithic *system* limit entails a myriad of assumptions, including the microarchitecture of the chip, the switching energy of a critical path gate, heat removal capacity of the packaging, and target clock frequency. For example, the fundamental limit on the signal energy E_S, which must be transferred during a binary switching transition, is given by the expression $E_S = (\ln 2)kT$, where k is Boltzmann's constant and T is absolute temperature [2]. Derivation of precisely this result based on two entirely distinct physical models confirms its validity. The first model is an ideal MOSFET operating in a CMOS inverter circuit at the limit of its capacity for binary signal discrimination. The second is an isolated interconnect treated as a noisy communication channel. This result receives further support from the observation that based on a Boltzmann distribution, the probability of error is 0.5 for a binary transition with signal energy transfer $E_s = (\ln 2)kT$. Clearly, E_s defines a fundamental limit since its value is independent of the properties of any material, device, or circuit used to implement the switching transition. However, evaluation of this fundamental limit still requires assumption of an operating temperature T.

The papers of this issue address the complete hierarchy of theoretical and practical limits on GSI. In the opening article of the issue, R. W. Keyes presents a comprehensive overview of fundamental limits in silicon technology. Limits dealing with power, energy, scaling, random dopant placement, and complexity as well as soft errors due to cosmic rays and alpha particles are discussed. The need for new transport models for both transistors and wires as their dimensions approach the mean free path of carriers is pointed out. In addition, Keyes provides experimental data illustrating the exponential decrease in the energy used to perform a logic operation that has occurred over the past six decades. It is interesting to observe that according to this data, the energy per logic operation for year 2000 exceeds the fundamental limit by approximately nine decades!

Material and process limits in silicon technology are addressed by J. D. Plummer and P. B. Griffin in the second article of the issue. The authors argue cogently that to maintain the pace projected by the 1999 International Technology Roadmap for Semiconductors (ITRS) over the next 15 years will require the introduction of new materials, processes, and device structures. MOSFET gate insulators with higher dielectric constants than silicon dioxide will be necessary. Metal gate electrodes will be needed to replace polycrystalline silicon and lower resistance source and drain regions as well as smaller contact resistance will be required. Although there are no apparent *fundamental* barriers to the introduction of these new materials and processes, they represent rather unprecedented departures from previous practice and therefore at the least can be expected to be quite disruptive.

D. J. Frank and colleagues then explore device scaling limits of single- and double-gate Si MOSFETs and their application dependencies in digital logic as well as static and dynamic memory circuits. Threshold voltage rolloff and subthreshold swing rollup due to short channel effects, tunneling current due to thin gate insulators, minimum signal swing

Publisher Item Identifier S 0018-9219(01)02075-8.

(or supply voltage in CMOS circuits) for binary signal discrimination, and the impact of random channel dopant atom placement are considered. Roughly speaking, the projections conclude that in bulk technology, minimum supply voltages in the 0.7–1.0-V range, minimum threshold voltages in the 0.2–0.5-V range, minimum equivalent gate insulator thickness in the 1.0–1.5-nm range, and minimum channel lengths in the 20-nm range are feasible. For double-gate technology, supply voltage, threshold voltage, and gate insulator thickness projections are comparable to those of bulk technology while somewhat smaller minimum channel lengths in the 15-nm range are anticipated *assuming* a suitable double gate manufacturing technology will be invented.

An eloquent personal perspective of B. Gilbert on the endless future opportunities for analog circuitry is the topic of the fourth paper of this issue. With emphasis on the role of bipolar technologies, the author points out that there are actually three fields of electronics today. The two major groupings are analog and digital and the third is a smaller but well-defined and rapidly growing group of techniques, which are called quasi-analog or binary analog, exemplified by the sigma–delta technique that represents signals by samples quantized only along the time axis. Digital signal processing and RF wireless communications systems illustrate impressively the simultaneously symbiotic and synergistic relationship between the three circuit groupings.

In the fifth contribution to this issue, J. A. Davis and his coauthors discuss the hierarchy of limits on GSI imposed by interconnects. Fundamental limits on interconnect latency and energy transfer per binary transition are defined by the speed of light and, in essence, by Shannon's classical theorem for the capacity of a communication channel, respectively. Severe surface scattering of electrons in extremely fine conductors caused by cross-sectional dimensions smaller than the mean free path of bulk electrons increases the effective resistivity of a conductor and represents a daunting interconnect materials challenge. This effective resistivity ρ increase coupled with normal cross-sectional area (F^2) reduction as minimum feature size (F) decreases cause drastic increases in resistance per unit length ($R/L = \rho/F^2$) of interconnects. The use of repeater circuits and the much more aggressive objective of three-dimensional (3-D) structures consisting of multiple levels of transistors and interconnects offer prospects for relief from the escalating tyranny of interconnects in GSI.

The fifth and highest level of the hierarchy of limits on GSI is the system level, whose salient characteristics are epitomized by the microprocessor. In the sixth paper of this issue, R. Ronen and colleagues discuss coming challenges in microarchitecture and architecture of microprocessors. In terms of the key performance metric of throughput measured in instructions executed per second, microprocessor performance has been doubling annually over the past two decades due to the combined effects of process technology, architecture, and design tool improvements. However, extremely formidable new challenges such as power dissipation, wire delays, and soft errors caused by alpha particles and secondary cosmic rays are now at hand. New architectures and microarchitectures can help address these challenges through higher levels of parallelism including simultaneous multithreading and single-chip multiprocessors to maintain the historic rate of improvement in microprocessor capabilities.

The central thesis of a hierarchy of fundamental, material, device, circuit, and system limits on GSI has proven to elucidate *physical* limits. In the seventh paper of this issue, R. Bryant and his coauthors provide an erudite treatise on the limitations and challenges of computer-aided *design* technology for CMOS GSI. Design technology is concerned with the automated conception, synthesis, verification, and testing of gigascale systems. Design technology faces intrinsic limitations inherent in the computational intractability of design optimizations involving practical tradeoffs among multiple objectives. Consequently, heuristic approaches must be ever present in design technology and therefore may represent its sole fundamental limit. Measurement and continuous improvement of design technology are needed in order to achieve the full potential of GSI. To this end, relevant and useful metrics of the design process must be identified and exercised.

Throughout the past three to four decades, optical lithography has been one of, if not, *the* key enabler of the semiconductor industry. In his paper "Limits of Lithography," L. Harriott examines the limits of optical lithography and possible future technologies from both a technical and economic perspective. The theoretical limit of resolution of optical lithography is one half the wavelength, which imposes a minimum feature size of approximately $157/2 = 78$ nm using a fluorine excimer laser photon source, calcium fluoride lenses, fluorine-doped fused silica masks, and as yet undeveloped photoresist material. Feature sizes smaller than approximately 78 nm will require nonoptical technologies such as extreme ultraviolet (EUV) and electron projection lithography (EPL). It is not entirely clear if and when each of these technologies will be used in semiconductor manufacturing.

The most powerful driver of semiconductor technology throughout its history has been a continuous manufacturing cost reduction of 30% per electronic function per year. In the ninth contribution to this issue, R. Doering and Y. Nishi suggest a novel methodology for study of a hierarchy of limits on integrated circuit manufacturing. The scope of this hierarchy includes the following aspects of the semiconductor landscape: process, equipment, factory, test/assembly, devices, circuits, and business/economics. The realm of semiconductor manufacturing *per se* includes the process, equipment, and factory domains and is responsive to the remaining domains, which are treated as the semiconductor product realm. Steady-state limits on stable single-product high-capacity factories are studied to establish a state-of-the-art manufacturing baseline that is projected to year 2014, the horizon of the 1999 International Technology Roadmap for Semiconductors. Cycle-time models project a vast potential for improvements in semiconductor manufacturing such as 30 times faster throughput intervals for 24-wafer lots compared with current volume production benchmarks.

In the tenth and final contribution to this issue, T. Ohmi and colleagues propose an aggressive new paradigm for silicon technology to implement future mixed-signal gigascale systems operating in the gigahertz frequency range. This future technology features: 1) a thin silicon device layer separated from a metal substrate by a buried silicon nitride insulating layer; 2) a MOSFET gate stack consisting of an interfacial layer of silicon nitride under a tantalum oxide layer under a tantalum gate metal electrode; and 3) a multilevel copper interconnect network with helium gas isolation. Silicon nitride thermal vias are inserted to prevent excessive increases in temperature of the multilevel interconnect network. Low-temperature low-energy microwave-excited plasma processing is used to achieve defect-free 3-D integration of multiple levels of thin-film transistors.

The key material, device, and circuit limits described in this issue may well be achieved within the first two decades of this century. **What lies beyond these limits?** Two hypothetical scenarios serve to reveal a broad range of possibilities. The first might be called *incremental* and the second *exponential*. Both are suggested by analogy.

Let us consider first the incremental scenario. The most important new structural material that came into wide use in the second half of the 19th century was steel. It gained importance throughout the first half of the 20th century, but its significance declined in the second half of the century chiefly due to increasing use of aluminum and a wide variety of synthetic materials such as plastics. The dominant electronic material of the second half of the 20th century was silicon. By analogy with steel, one might project that silicon will continue to serve as the dominant material of the semiconductor and, hence, the electronics industry for at least the first half of the 21st century!

Now let us explore the exponential scenario. Technological historians have observed on numerous occasions that commercially successful technologies tend to follow a characteristic "S-curve" pattern of development if the state-of-the-art (SoA) is plotted on a vertical axis and calendar year (Y) on the corresponding horizontal axis. Initially, at the bottom of the S-curve, the SoA advances rather slowly as the rudiments of the technology are explored in academic or industrial laboratories. When the commercial potential of the technology is recognized, relatively large investments are made resulting in a rapid advance of the SoA until physical or economic limits cause a leveling off or saturation at the top of the S-curve.

During the first half of the 20th century, the advance of vacuum tube electronics proceeded along an S-curve trajectory. Then, in midcentury, a very rare discontinuity occurred. The advance of electronics effectively jumped the gap from the vacuum tube S-curve to a second S-curve representing semiconductor technology. As the papers of this issue forecast, a saturation of the semiconductor S-curve is expected within two decades. Again by analogy, what may lie beyond saturation of the semiconductor S-curve is a virtually unprecedented second jump to a third S-curve, possibly implemented by quantum nanotechnology, which will continue for many decades to come. Fig. 1 illustrates this "triple S-curve" exponential scenario.

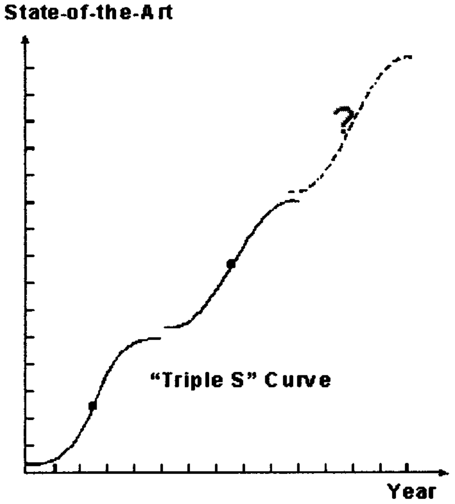

Fig. 1 State-of-the-art versus calendar year.

In conclusion, *analogical* limits suggest that under either the incremental or the exponential scenario, the impact of electronics technology on our lives will continue to grow during the 21st century.

James Meindl
Guest Editor
Georgia Institute of Technology
Atlanta, GA 30332-0269 USA

References

[1] J. D. Meindl, "Low power microelectronics: Retrospect and prospect," *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.

[2] J. D. Meindl and J. A. Davis, "The fundamental limit on binary switching energy for terascale integration (TSI)," *IEEE J. Solid-State Circuits*, vol. 35, pp. 1515–1516, Oct. 2000.

James Meindl (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Carnegie Mellon University, Pittsburgh, PA.

He is the Director of the Joseph M. Pettit Microelectronics Research Center and has been the Joseph M. Pettit Chair Professor of Microelectronics at Georgia Institute of Technology since 1993. He was Senior Vice President for Academic Affairs and Provost of Rensselaer Polytechnic Institute from 1986 to 1993. He was with Stanford University from 1967 to 1986 as the John M. Fluke Professor of Electrical Engineering, Associate Dean for Research in the School of Engineering, Director of the Center for Integrated Systems, Director of the Electronics Laboratories and Founding Director of the Integrated Circuits Laboratory.

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 a Benjamin Garver Lamme Medal from ASEE, an IEEE Education Medal, an IEEE Solid-State Circuits Medal, and an IEEE Beatrice K. Winner Award. He has also been awarded the IEEE Electron Devices Society's J. J. Ebers Award, the Hamerschlag Distinguished Alumnus Award, Carnegie Mellon University, the 1999 SIA University Research Award, and the IEEE Third Millennium Medal.