

CMOS Nanometer scaling, Limited!

Dr. Danny Rittman
July 2006

Introduction

Nanometer design complexity has become a serious issue in nanometer designs. The main reason for the constant growing complexity is the exponential rise in the number of devices integrated in a single chip and technology scaling. As the process downscale into the deep nanometer range many new issues arise while old ones become sever. Phenomenon like reliability, noise, power and lithography are well known for decades, hence they become serious limiting factors when it comes to deep nanometer scaling. As a direct result the silicon capacity is worsen and the design productivity is getting lower. The steady downscaling of transistor dimensions over the past decades has been the main incentive to the growth of microchips. The more an IC is scaled, the higher becomes its packing density, the higher its circuit speed, and the lower its power dissipation. These have been the key in the evolutionary progress leading to today's computers and communication systems that offer superior performance, dramatically reduced cost per function, and much-reduced physical size compared to their predecessors. Today the VLSI technology comprises CMOS devices because of their unique characteristic of negligible standby power, which allows the integration of hundreds of millions of transistors on a chip with only a very small fraction of them switching at any given instant. However, as the CMOS dimension is scaled to the deep nanometer arena, in particular the channel length, the electrical barriers in the device begin to lose their insulating properties due to thermal injection and quantum-mechanical tunneling. This results in a rapid rise of the standby power of the chip, placing a limit on the integration level as well as on the switching speed. The major limiting factors are power and threshold voltage, tunneling leakage through gate oxide, lithography, short-channel effect, high-field effects, dopant number fluctuations, interconnect delays and electrostatic scale length. (See Figure #1) Both the standby power and the active power of a chip will increase precipitously below the 45nm technology generation. To extend CMOS scaling to the shortest channel length possible while still gaining significant performance benefit, the industry is researching throughout the horizon, looking for new design concepts. One of them is an optimized, vertically and laterally non-uniform doping design. (Also called Super-Halo) It is anticipated that room-temperature CMOS will be scaled to 20nm channel length with the super-halo profile. Low-temperature CMOS allows additional design space to further extend CMOS scaling to near 10nm. At this point the industry will reach its critical barrier. The big question is where do we go from here?

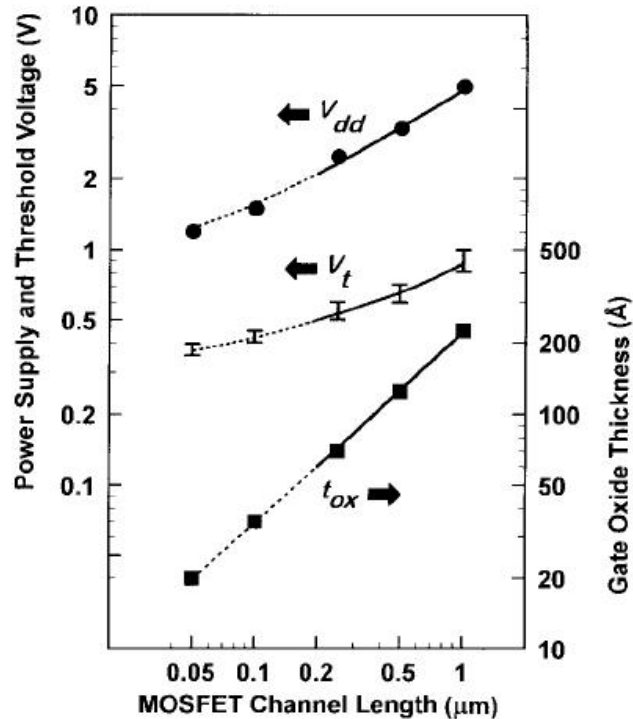


Figure 1: Power supply voltage (V_{dd}), Threshold voltage (V_t) and gate oxide thickness (t_{ox}) vs. CMOS channel length

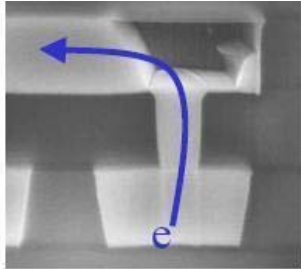
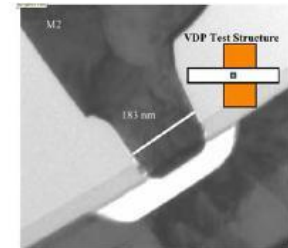
CMOS scaling – The basics

As chip designers are constantly downscaling CMOS the voltage level and the gate-oxide thickness are also reduced. Since the electron thermal voltage is a constant for room-temperature electronics, the ratio between the operating voltage and the thermal voltage inevitably shrinks. This leads to higher source-drain leakage currents stemming from the thermal diffusion of electrons. At the same time, the gate oxide has been scaled to a thickness of only a few atomic layers, where quantum-mechanical tunneling gives rise to a sharp increase in gate leakage currents. Most IC design problems can be formulated as a multi-objective constrained optimization problem. The optimality and scalability of the optimization engine significantly impact the design methodology and quality. Furthermore known nanometer phenomenon like reliability, power and noise sensitivity are becoming a significant limiting factor. (See Figure #2)

Via reliability: Limits to scaling

Smaller features >> larger stress gradients

Thinner barriers >> less coverage >> poor Cu/barrier adhesion



Electromigration: Limits to scaling

Smaller features >> less material to migrate

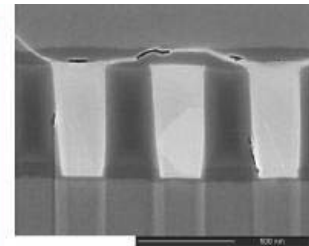
Thinner barriers >> less barrier coverage

Increasing product requirements

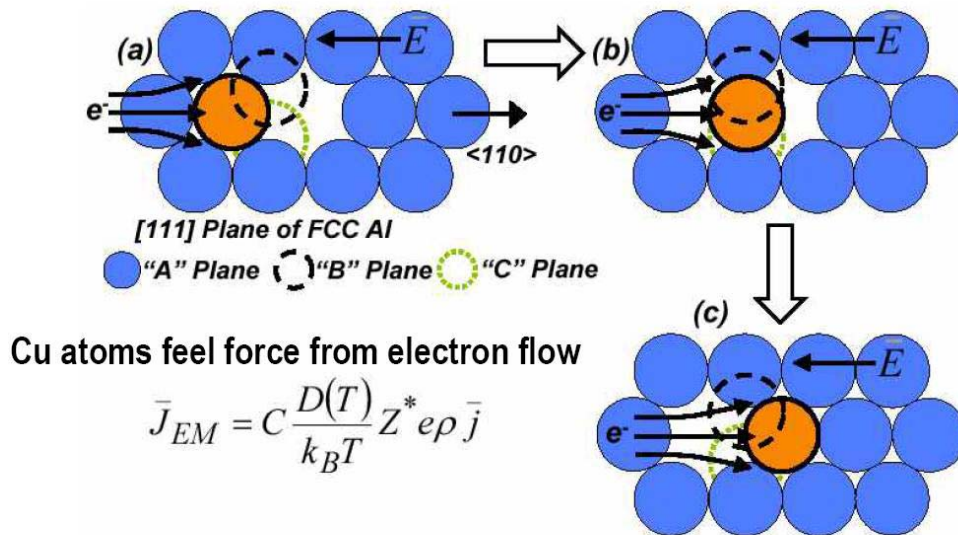
Dielectric Reliability: Limits to scaling

Low k inter-level dielectrics: soft / poor adhesion

Low k dielectric barriers: Less hermetic



Mechanism of Electromigration



Ref: (1) Electronic Thin Film Science, K.-N. Tu, J. W. Mayer, L. C. Feldman, 1992.

Figure #2: Scaling Reliability issues
Images Source: Novellus

Power supply & threshold voltage V_{th}

Design's power supply has been one of the major concerns for the past decade. One of the fringe benefits of scaling is power supply reduction. The power-supply voltage has not been decreased at a proportional rate to the channel length. This means that the field has been gradually rising over the generations between $1\mu\text{m}$ and $0.1\mu\text{m}$ channel lengths. Luckily, thinner oxides are more reliable at high fields, therefore allowing operation at the reduced but non scaled supply voltages. Below $0.1\mu\text{m}$, threshold voltage deviates even further from the past scaling behavior. MOSFET threshold voltage is defined as the gate voltage at which significant current begins to flow from the source to the drain. Below the threshold voltage, the current does not drop immediately to zero. Rather, it decreases exponentially, with a slope on the logarithmic scale inversely proportional to the thermal energy kT . This is because some of the thermally distributed electrons at the source of the transistor have high enough energy to overcome the potential barrier controlled by the gate voltage and flow to the drain. Such a sub-threshold behavior follows directly from fundamental thermodynamics and is independent of power-supply voltage and channel length. The active power of today's high-end microprocessors is already in the 50–100W range.

Barring a major breakthrough in power management systems via architectural innovation, expensive packaging solutions will very soon be required in order to dissipate the heat generated by the chip. There are other schemes for meeting leakage power requirements. For example, one can fabricate multiple threshold-voltage devices on a chip. Low-threshold devices could be used in critical logic paths for speed, while high-threshold devices would be used everywhere else, including memory arrays, for low standby power. One can also sense the circuit activity and cut off the power supply to logic blocks that are not switching—an approach known as sleep mode. Other possibilities include dynamic threshold devices, for which the threshold voltage is controlled by a back-gate bias voltage in either bulk or silicon-on-insulator device structures. Yet another option is low-temperature CMOS. Low-temperature operation not only worsens the sub threshold slope and improves mobility, but also reduces wire resistance. However, all of these solutions generally carry a cost in density and complexity. In today's deep nanometer designs this is a high price that can not be afforded at all times.

Gate-oxide tunneling

In order to keep short-channel effect under control and to maintain a good sub threshold turn-off slope, gate oxide thickness is reduced nearly in proportion to channel length. This is necessary in order for the gate to retain more control over the channel than the drain. This thickness comprises only a few layers of atoms and is approaching fundamental limits. While it is amazing that SiO_2 can carry us this far without being limited by extrinsic factors such as defect density, surface roughness, or large-scale thickness

and uniformity control, oxide films this thin are subject to quantum-mechanical tunneling, giving rise to a gate leakage current that increases exponentially as the oxide thickness is scaled down. In the direct-tunneling regime, the current is rather insensitive to the applied voltage or field across the oxide, so reduced-voltage operation will not buy much relief. Although the gate leakage current may be at a level that is negligible compared with the on-state current of a device, it will first have an effect on the chip standby power. Note that the leakage power will be dominated by turned-on n-MOSFETs, in which electrons tunnel from the silicon inversion layer to the positively biased gate. Edge tunneling in the gate-to-drain overlap region of turned-off devices should not be a fundamental issue, since one can always build up the corner oxide thickness by additional oxidation of polysilicon after gate patterning. P-MOSFETs have a much lower leakage than n-MOSFETs because there are very few electrons in the P+ polysilicon ("poly") gate available for tunneling to the substrate, and hole tunneling has a much lower probability. Dynamic memory devices have a more stringent leakage requirement and therefore must impose a higher limit on gate-oxide thickness. Another issue with the thin gate oxide is the loss of inversion charge and therefore trans-conductance due to inversion-layer quantization and polysilicon-gate depletion effects. Quantum mechanics dictates that the density of inversion electrons peaks at approximately 1nm below the silicon surface, which effectively reduces the gate capacitance and therefore the inversion charge to those of an equivalent oxide $\sim 0.4\text{nm}$ thicker than the physical oxide. Similarly, depletion effects occur in polysilicon in the form of a thin space-charge layer near the oxide interface which acts to reduce the gate capacitance and inversion-charge density for a given gate drive. The percentage of gate-capacitance attenuation becomes more significant as the oxide thickness is scaled down. CMOS design space is severely constrained by voltage and oxide limits below 45nm dimensions.

Lithography

It has been the ability of lithographic patterning to continually reduce device features in the lateral dimensions that has directly led us into the ultra-large scale integration (ULSI) era. Extensions beyond today's lithography technologies are required in order to bring CMOS into the nanoscale regime. Optical photolithography has exceeded previously predicted resolution limits many times over by a combination of improved lenses with higher numerical aperture and the use of shorter wavelength illumination. RET techniques relies on material or topographic changes on the optical mask to vary the phase of the illuminating radiation. The resulting interference effectively sharpens the image at the wafer plane. Because this effect is geometry dependent, phase-shifting technique has thus far not been demonstrated to be generally applicable to arbitrary device geometries which may be encountered in a chip design. With the exception of near-field techniques, which may be impractical for device fabrication applications, there are no current expectations that optical lithographic techniques will extend into the sub 45nm regime. For the fabrication of such ultrasmall devices, optical

lithography may be used for non critical levels, in a mix-and-match scheme, where the critical features are defined by electron beam lithography or X-ray lithography. X-ray lithography is a prime candidate for high-resolution patterning of critical features for sub 45nm CMOS applications. The challenges in implementing an X-ray lithography technology lie primarily in mask fabrication. X-ray masks consist of thin membranes of Si or a Si compound such as Si₃N₄ or SiC, patterned with an X-ray absorbing material. Electroplated gold has been used as the absorber material. Precise control of mechanical stress in the absorber-covered membrane must be maintained, as it has a direct effect on the image placement accuracy in X-ray lithography. In addition, since X-ray proximity printing is a replication, more stringent control of defects is required as compared to systems using image reducing optics. Another form of X-ray lithography under investigation for sub 45nm applications has come to be known as extreme ultraviolet (EUV) lithography. This technique uses reflective optics with a 4 reduction scheme. Key challenges to this technology lie in the radiation source, in the multilayer thin-film mirror optics, and in mask fabrication.

Electron beam lithography has been the lithography workhorse in nanostructure patterning within the research environment for years. Since Gaussian probes can be focused to a spot of only a few nanometers, the resolution limits of e-beam lithography are primarily determined by beam-resist-substrate interactions and are thought to be 10nm for conventional lithography and resist systems. Thus, for Gaussian beam systems, resolution limitations are generally not a consideration in CMOS device applications. Pattern distortion due to electron scattering, known as proximity effects, can be effectively compensated for by a variety of correction schemes. High-accuracy pattern placement can also be achieved with good stage interferometry and control over noise in the electron beam deflection system. The key challenge for e-beam lithography lies in throughput. Even with mix-and-match lithography schemes, Gaussian beam systems are incapable of patterning the ten or more pixels required for product level chips in a reasonable time. Shaped beam systems and character projection systems offer several orders of magnitude improvement in speed by patterning many pixels in parallel, however the increase in beam current needed to achieve high throughput will limit the achievable resolution because of Coulomb interactions and off-axis aberrations. An alternative technique which has received recent interest is electron beam projection lithography. Proposed projection systems use reduction optics with a mask patterned alternately with an electron absorbing material on an electron transparent or electron scattering substrate. Electron optical models predict that such systems can image up to ten pixels per exposure field with a resolution near 50nm, limited by Coulomb interactions and geometric aberrations. Recently, there has been significant activity in patterning of thin films with proximal probes. Typically, the scanning speeds of such systems are quite low (1 mm/s), so that significant throughput improvements will have to be realized before scanning probes can serve as the basis for lithography for CMOS applications.

Short-Channel Effect

Short-channel effect is the decrease of threshold voltage in short-channel devices due to two-dimensional (2D) electrostatic charge sharing between the gate and the source drain regions. Short-channel effect plays a key role in V_t -tolerances which determine the minimum acceptable V_t . To scale down MOSFET channel length without excessive short-channel effect, both the oxide thickness and the gate-controlled depletion width in silicon must be reduced in proportion to L . The latter requires increased channel doping concentration which, for a uniformly doped channel, leads to higher depletion charge and electric field at the silicon surface. These in turn cause the potential across the oxide and therefore the threshold voltage to go up. To reduce the gate-controlled depletion width while fulfilling the V_t reduction trend depicted, a low-high, channel doping is needed below channel length. For the same gate depletion width, the surface electric field and the total depletion charge of an extreme retrograde channel is one-half of that of a uniformly doped channel. This reduces threshold voltage and improves mobility. Retrograde channel doping represents a vertically non-uniform profile that allows the threshold voltage to be decoupled from the gate-controlled depletion width. However, the body-effect coefficient and the inverse sub-threshold slope are still coupled to the gate depletion width W_{dm} . For a given t_{ox} , reduction in improves short-channel effect but compromises substrate sensitivity and sub-threshold slope. Halo doping or non-uniform channel profile in the lateral direction provides yet another degree of freedom which can be tailored to further minimize V_t tolerances due to short-channel effect. In an idealized, 2-D non-uniform channel-doping profile, pockets of high-doping regions are placed at two lower corners of the gate controlled depletion region where the potential difference (band bending) between the source/drain and the substrate is the highest. These regions are partly depleted by the source/drain fields and shield the rest of the channel region from further penetration. In the middle of the channel and under the gate, the doping concentration is low to keep the gate depletion charge to a minimum. Such a 2-D doping profile can be formed with self-aligned halo implants made together with the source-drain implant. Halo doping has been shown to significantly improve short-channel effect in deep nanometer range.

Discrete doping effects

Another physical effect that may limit scaling is the discreteness of the dopant atoms. Although the average concentration of doping is quite well controlled by the standard ion implantation and annealing processes, these processes do not control exactly where each dopant ends up. Consequently there is randomness at the atomic scale, resulting in spatial fluctuations in the local doping concentration, and these in turn cause device-to-device variation in MOSFET threshold voltages. As MOSFET technology nears the end of scaling, it will be readily possible to make devices with fewer than 100 dopant atoms controlling the threshold voltage. Since fluctuations in dopant

number have a standard deviation equal to the square root of the number of dopants, in keeping with Poisson statistics, threshold variation may very well become quite large, making the design of robust circuits very difficult.

Scaling-Related Problems & Solutions

With the move advanced into nanometer technologies scaling-related issues become key factors. These issues include noise, reliability, power, interconnect delays and thermal limitations. These problems have significantly complicated the design process. In order to provide efficient solutions the industry is moving in several directions. One direction is to develop novel predictable and robust synthesis techniques. In order to reduce the uncertainty of interconnect delays, for example, gain-based-synthesis and physical-synthesis approaches have been proposed and successfully used in the existing synthesis tools. To tolerate interconnect latency, techniques for multi-cycle on-chip communication also have been proposed and showed promising results. These techniques are based on the regularly distributed register architectures and latency-insensitive designs. They greatly reduce the difficulty in handling interconnect uncertainty in high-level designs. A second approach is to develop integrated modeling, analysis, and synthesis capabilities. For noise and power control, for example, interconnect capacitance needs to be modeled and estimated throughout the synthesis process (i.e., during logic optimization, technology mapping, placement, and routing). The models and analysis tools used at different stages should be consistent and have increasing accuracy as more physical information is available. In addition, various optimization operations need to be applied throughout the design process from the logic domain to the physical domain and guided by the analysis results of increasing accuracy. These operations include netlist remapping, driver sizing, buffer insertion, and wire spacing for noise and power optimization. A unified tool (single binary) with integrated synthesis, physical-design, and analysis capabilities provides a promising solution. The industry must attack the subject from both aspects the higher level design (modeling, RTL) and backend. (Physical Layout) Only a combined, integrated and synchronized approach will hold a deep nanometer range scaled design.

Conclusion

As CMOS devices will continue to downscale into the deep nanometer range with improved device performance and lower power, it will be running into fundamental barriers of physics. Difficult challenges lie ahead in tightening process tolerances to satisfy more stringent defect density and reliability requirements in future generation CMOS technologies. Some of the solutions call for a paradigm shift and costly buildup of new infrastructures, for example in the case of X-ray lithography. Others require near atomic-level thickness control and nanometer-scale lateral-dimension inspection and control. Much lower defect densities and higher device yields than today's standard will undoubtedly be required when multi-billion transistors are

fabricated on a single chip. Scaling below 45nm channel length faces several fundamental limiting factors stemming from electron thermal energy and quantum-mechanical tunneling. Many of the potential barriers in a MOSFET that kept the standby leakage low are losing their effectiveness when scaled to lower barrier heights or thinner widths. Certainly, both the standby power and the active power of a high-performance processors and chips will rise. As a direct result, the performance gained from scaling will slow. Yet, by using properly optimized doping profiles and pushing the silicon depletion width to the tunneling limit, it is likely that mainstream CMOS scaling will be extended to 20nm channel length with non scaled gate oxides and voltage levels. Downscaling further, cooling to low temperature might provide the additional design space needed to extend CMOS devices to 10nm for server applications. The million dollar question is what's next? Well, some experts claim that this is probably the end of CMOS as we know it. Either the industry will break the atomic barrier, finding the way to utilize the technology further more or we are bound to make the next breakthrough in VLSI. Like the move from the electric tube to the transistor and the move from the transistor to the integrated circuit. No doubt, a VLSI breakthrough will be needed soon.

References:

1. International Technology Roadmap for Semiconductors, 2004 Edition; <http://www.itrs.net/Common/2004Update/2004Update.htm>.
2. Taurus-Process & Taurus-Device User's Manuals V-2003.12, Synopsys, 2003.
3. R. Lindsay, et al., "A Comparison of Spike, Flash, SPER and Laser Annealing for 45nm CMOS," Proc. MRS Symp., Vol. 765, p. D.7.4, 2003.
4. K. Mistry, et al., "Delaying Forever: Uniaxial Strain Silicon Transistors in a 90nm CMOS Technology," Proc. VLSI Research Symp., pp. 50-51, 2004.
5. V. Moroz, et al., "Analyzing Strained-silicon Options for Stress-Engineering Transistors," Solid State Technology, p. 49, July 2004.
6. C.S. Smith, "Piezoresistance Effect in Germanium and Silicon," Phys. Rev., Vol. 94, No. 1, pp. 42-49, 1954.
7. S. Inaba, et al., "High Performance 35nm Gate Length CMOS with NO Oxynitride Gate Dielectric and Ni Salicide," Proc. IEDM, pp. 641-644, 2001.
8. S. Pidin, et al., "A Novel Strain-enhanced CMOS Architecture Using Selectively Deposited High-tensile and High-compressive Silicon Nitride Films," Proc. IEDM, pp. 213-216, 2004.
9. V. Moroz, X. Xu, "Exploring Stress Engineering Approaches for the 45nm Technology Node," to be presented at Electrochemical Soc. Mtg., May 2005.
10. S.E. Thompson, et al., "A Logic Nanotechnology Featuring Strained Silicon," IEEE Electron Dev. Lett., Vol. 25, No. 4, pp. 191-193, 2004.
11. F. Nouri, et al., "A Systematic Study of Trade-offs in Engineering a Locally Strained pMOSFET," Proc. IEDM, pp. 1055-1058, 2004.
12. J.G. Fossum, et al., "Pragmatic Design of Nanoscale Multigate CMOS," Proc. IEDM, pp. 613-616, 2004.

13. I.J. Malik, et al., "Optoelectronic Substrates by SiGen NanoTec — A General Layer-transfer Approach," 2004 Elec. Chem. Proc., Vol. 2004-07, pp. 543-554.
 14. I.J. Malik, et al., "The Genesis A General Layer-transfer Method for Electronic Applications," Spring 1999 MRS Symp. Tech. Proc., 1999.
 15. I.J. Malik, et al., "Fully-Integrated Plasma-activated Bonding (PAB) for High Volume SOI Substrate Mfg. Proc.," Spring 2003 ECS Mtg., Ext. Abs., 2003.
 16. A. Thilderkvist, et al., "Surface Finishing of Cleaved SOI Films Using EPI Technologies," Proc. IEEE International SOI Conference, pp. 12-13, 2000.
 17. Z. Krivokapic, et al., "Locally Strained Ultra-thin Channel 25nm Narrow FDSOI Devices with Metal Gate and Mesa Isolation," Proc. IEDM, pp. 445-448, 2003.
 18. S.E. Thompson, et al., "Key Differences for Process-induced Uniaxial vs. Substrate-induced Biaxial Stressed Si and Ge Channel MOSFETs," Proc. IEDM, pp. 221-224, 2004.
 19. H.K. Kirk, et al., "Wafer-level Uniaxially Strained s-SOI by Direct Mechanical Stress," Proc. IEEE International SOI Conference, pp. 102-103, 2004.
 20. T. Mizuno, et al., "Physical Mechanism for High Hole Mobility in (110)-Surface Strained- and Unstrained-MOSFETs," Proc. IEDM, pp. 809-812, 2003.
 21. J.R. Hwang, et al., "Symmetrical 45nm PMOS on (110) Substrate with Excellent S/D Extension Distribution and Mobility Enhancement," Proc. Symp. on VLSI Tech., pp. 90-91, 2004.
 22. M. Yang, et al., "High Performance CMOS Fabricated on Hybrid Substrate with Different Crystal Orientations," Proc. IEDM, pp. 453-456, 2003.
 23. Z. Krivokapic, et al., "Strain Relaxation in Narrow Width Strained Silicon Devices with Poly and Metal Gates," 2004 Elec. Chem. Proc., Vol. 2004-07, pp. 459-469.
- Victor Moroz is a principal engineer in TCAD at Synopsys, 700 East Middlefield Rd., Mountain View, CA 94043; ph 650/584-5458. Dipankar Pramanik is group director, TCAD DFM solutions, at Synopsys. Francois Henley is president and CEO at Silicon Genesis Corp. Philip Ong is VP of engineering at Silicon Genesis Corp.
24. Chang, C.C., Cong, J., and Xie, M., "Optimality and Scalability Study of Existing Placement Algorithms," Asia Pacific Design Automation Conference, Jan. 2003.
 25. Chan, T., Cong, J., Shinnerl, J., and Sze, K., "An Enhanced Multilevel Algorithm for Circuit Placement," International Conference on Computer-Aided Design, Nov. 2003.
 26. Karypis, G., et. al., "Multilevel Hypergraph Partitioning: Application in VLSI Domain," Design Automation Conference, June 1997.
 27. Cong, J., Xie, M., and Zhang, Y., "An Enhanced Multilevel Routing System," International Conference on Computer-Aided Design, Nov. 2002.
 28. Cong, J., Fan, Y., Han, G., Yang, X., and Zhang, Z., "Architecture and Synthesis for On-Chip Multicycle Communication," IEEE Trans. on CAD, Vol. 23, April 2004.