Localized Heating Effects and Scaling of Sub-0.18 Micron CMOS Devices

Eric Pop, Kaustav Banerjee, Per Sverdrup*, Robert Dutton and Kenneth Goodson**

Department of Electrical (**and Mechanical) Engineering, Stanford University
*now at Intel Corp., 2200 Mission College Blvd, Santa Clara, CA 95052
Contact: epop@stanford.edu, Bldg 500 Room 501S, Stanford CA 94305-3030, fax 650.723.7657

ABSTRACT

We explore the generation and effect of phonon hot spots in silicon CMOS devices under steady state operation. The phonon Boltzmann Transport Equation (BTE) is used to extract generated phonon distributions for devices with channel length (L_{eff}) down to 90 nm. Estimates are made of the impact of phonon hot spots on transistor operation into the L_{eff} range approaching 10 nm. In this scaling limit the dimensions of the phonon hot spot are comparable to the device channel length. It is shown that localized drain region hot spots alter drain characteristics and, in the extreme scaling limit, may affect the resistance and electron injection at the source end, hence the current drive of a device. This is the first study that attempts to quantify non-equilibrium hot phonon effects in ultra-scaled CMOS devices and their implications for future scaling.

INTRODUCTION

Traditional studies of present and future scaling of CMOS devices focus on the electrical aspects of the problem — hot electron effects, oxide leakage, source/drain tunneling, polysilicon gate depletion, threshold voltage shifts due to short channel effects or inversion layer quantization — yet assume the problem to be isothermal. In modern and future ultra-scaled devices, large electric fields and very small dimensions produce hot electrons which are substantially away from thermal equilibrium with the lattice. The hot electrons scatter with and transfer their energy to the silicon lattice phonons before exiting through the contacts. Consequently, the phonons heat up and affect electron transport and device behavior. Fig. 1 shows a diagram of the energy transfer processes in silicon. Full understanding of the operation of a modern transistor cannot be accomplished by assuming an isothermal problem.

BTE SIMULATIONS

We use a phonon BTE solver [1] to study the lattice heating. The classical heat diffusion equation

$$C_s \frac{\partial T}{\partial t} = \nabla \cdot (k_s \nabla T) + Q''' \qquad (1)$$

(where C_s is the heat capacity per unit volume, k_s the thermal conductivity of silicon and Q''' the volumetric heat generation rate) underpredicts the peak temperature rise when applied at length scales less than the phonon-phonon mean free path Λ (about 300 nm in silicon at room temperature [2]). Of particular interest is the small region (a few tens of nm) just inside the device drain where most electron-phonon scattering takes place (see Fig. 2). Owing to its small dimensions (compared to Λ) and to the ballistic nature of phonon emission from this "hot spot," a reduced number of collisions in its vicinity prevent thermal equilibrium from being established between the hot spot and its surroundings. Consequently, the local hot spot temperature rises beyond diffusion theory predictions [3].

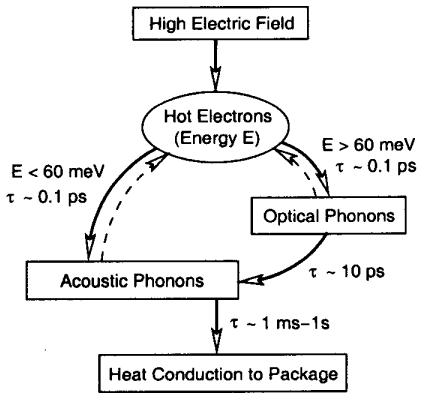

Fig. 1. Diagram and characteristic time scales of the energy transfer processes in silicon. Scattering with low group velocity longitudinal optical (LO) phonons is the dominant relaxation mechanism for electron energies above 60 meV. This creates a phonon energy bottleneck until the LO phonons decay into the faster acoustic modes.

Electron-phonon scattering selection rules [4] indicate that electrons with energies below 60 meV undergo acoustic phonon scattering, particularly with the longitudinal (LA) branch. The acoustic phonons have large group velocities (7000-8000 m/s) and quickly transport heat out of the region of most intense scattering. Hot electrons with energies above 60 meV participate in intervalley scattering most effectively with the 730 K longitudinal optical (LO) phonons in silicon (see [5] and Fig. 3). The energetic LO phonons have small group velocities (1000-2000 m/s) and are relatively stationary in the hot spot region. They anharmonically decay into the faster acoustic phonons which in turn transport the energy out.

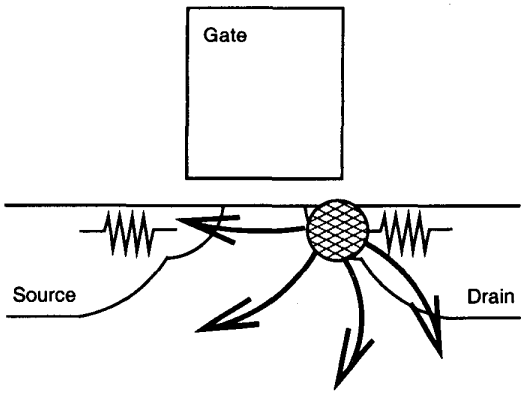

Fig. 2. Phonon hot spot in ultra-scaled MOSFET drain and its position relative to the device geometry. Hot phonons in and escaping from this region may impact device operation.

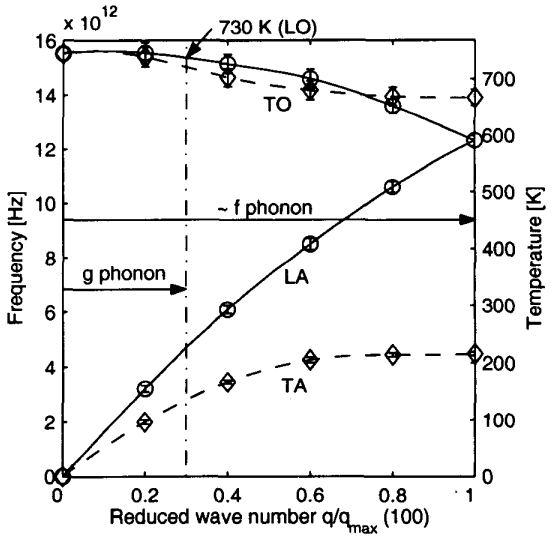

Fig. 3. Phonon dispersion relationship in silicon along the (100) direction, based on neutron scattering data [7]. The 730 K LO phonon dominates scattering for hot electrons with energies above 60 meV. The f and g phonons are involved in the intervalley scattering of electrons [4].

As the electron-LO phonon scattering time constant (about 0.1 ps) is much shorter than the LO-acoustic phonon decay time (about 10 ps), a phonon energy bottleneck is created and the hot spot temperature rises beyond diffusion theory predictions [6] as the hot optical phonons accumulate.

Our phonon BTE solver simplifies the phonon dispersion relationship (Fig. 3) by assuming a two-fluid phonon transport model [2] involving: a stationary optical reservoir mode and a propagating acoustic mode. The transformed phonon BTE in the relaxation time approximation

$$\frac{\partial e''}{\partial t} + \vec{v} \cdot \nabla e'' = \frac{e''_{eq} - e''}{\tau_{ph}} + Q''' \qquad (2)$$

(where e'' is the phonon energy per unit volume, v the phonon velocity and τ_{ph} the phonon energy relaxation time) is solved for the propagating mode and an integrated energy balance equation is solved for the reservoir mode (which is assumed to be stationary). Q''' is the volumetric heat generation rate absorbed from hot electrons which can be extracted from a device simulator. The individual mode temperatures are then computed. The resulting lattice temperature is obtained as an average of the mode temperatures, each weighted by its fractional contribution to the heat capacity.

Fig. 4 shows the two-dimensional lattice temperature profile computed using the BTE in a modern 180 nm n-channel CMOS device under steady state operation in the saturation regime. Figs. 5 and 6 show cross-sections of the lattice temperature along and perpendicular to the MOSFET channel, respectively, and how they compare with diffusion theory predictions. A large fraction of the phonon hot spot occurs just inside the device drain, within a few electron energy relaxation lengths from the region of highest electric field. The classical diffusion theory underestimates the temperature rise because it cannot account for the non-equilibrium condition between the two phonon modes. On the other hand, the same non-equilibrium situation makes it hard to define a "mode temperature," hence one must carefully interpret the BTE predictions. The BTE temperature profiles computed in Figs. 4 and 6 are therefore interpreted as indications of the local phonon energy density rather than actual, measurable quantities. The distribution along the channel from Fig. 5 was averaged over the channel depth to illustrate the sort of "average" lattice temperature that an electron (whose wavefunction spans several lattice sites) may experience as it travels from source to drain.

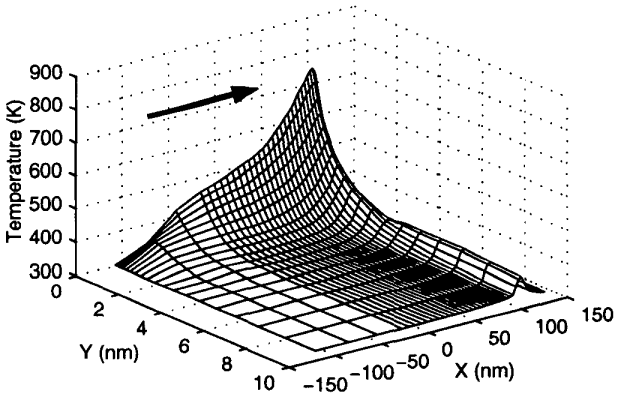

Fig. 4. Two-dimensional temperature profile computed by BTE solver for a device with L_{eff} = 180 nm under steady-state operation (V_{dd} = 1.8 V). The arrow indicates the direction of electron flow, from source to drain.

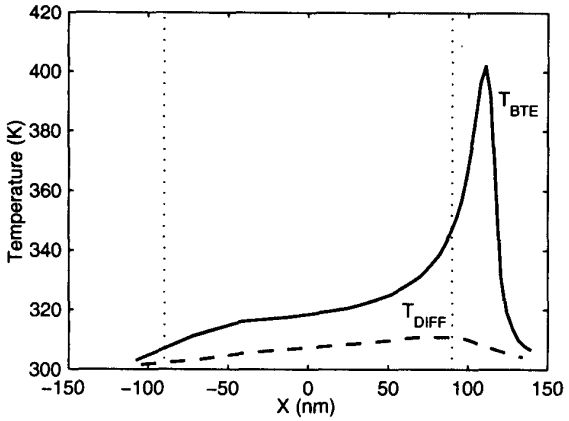

Fig. 5. Temperature distribution along the channel of a 180 nm MOSFET. The solid line represents the BTE solution averaged over the channel depth and the dashed line is the diffusion theory prediction. The vertical dotted lines represent the metallurgical source and drain junctions, respectively.

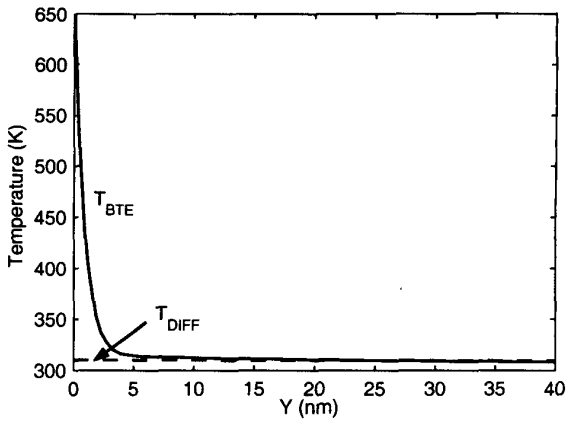

Fig. 6. Vertical temperature profiles through the 180 nm MOSFET at the drain metallurgical junction. The solid and dashed lines compare the BTE and diffusion equation predictions. The Si/SiO_2 interface corresponds to Y=0.

Hot Spot Scaling

As conventional CMOS devices are scaled down to nanometer dimensions, the characteristic phonon hot spot region near the drain does not scale proportionately. The size of the phonon hot spot is on the order of magnitude of the high electric field region near the drain in a modern 180 nm device. Near the ultimate scaling limit (L_{eff}=10-20 nm), the hot spot would be limited in size to the electron and LO phonon relaxation lengths, both on the order of 10 nm in silicon, and hence comparable to L_{eff}. BTE simulations show that for a modern 180 nm device the hot spot region, defined as the area under the peak (see Fig. 5) covering half of the temperature profile, is about 40 nm long and 10 nm deep. Fig. 7 shows the estimated scaling of the phonon hot spot with device channel length, down to the L_{eff} = 35 nm ITRS node where hot spot dimensions are already expected to be on the order of the electron and LO phonon relaxation length. The circles linked with solid lines in Fig. 7 come from simulation results performed on "well-behaved" devices of the respective technology nodes, whereas the diamonds represent our projections.

Despite the decreasing supply voltage of future technologies as proposed by the 1999 ITRS road map, simulations show that shrinking device dimensions lead to increased volumetric power densities. The average lattice temperature rise within a hot spot is proportional to this input power density, and a simple formula for estimating it has been theoretically proposed and experimentally supported [6]:

$$\Delta T = \frac{Q'\Lambda^2}{3A_{eff}k_s} \qquad (3)$$

where Q' is the input power per device width (the $I \cdot V$ product, I being the current per unit width), Λ is the acoustic phonon mean free path, k_s the thermal conductivity of bulk silicon and A_{eff} the effective hot spot area. The formula can be interpreted as the result of an additional thermal resistance which should be added to the diffusion theory predictions to account for the reduced scattering in the vicinity of a hot spot smaller than the phonon mean free path. This expression is used to estimate the average temperature rise inside phonon hot spots of future generations of transistors and the results are shown in Fig. 8. The error bars are due to the allowed tolerances in V_{dd} and device dimensions from the ITRS road map.

Outlook

Challenges of scaling device L_{eff} into the range approaching 10 nm pose many open questions; the following observations are but a first attempt to quantify the contribution of sub-

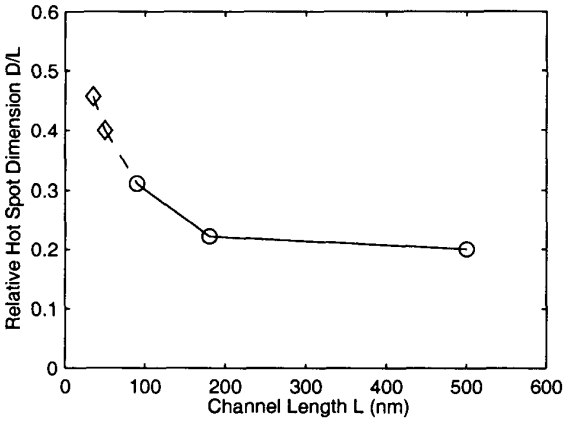

Fig. 7. Ratio of characteristic hot spot dimension (D, along device channel) to channel length (L) along the ITRS road map. At shorter channel lengths the region of high density LO phonons is expected to be comparable in size to the channel length.

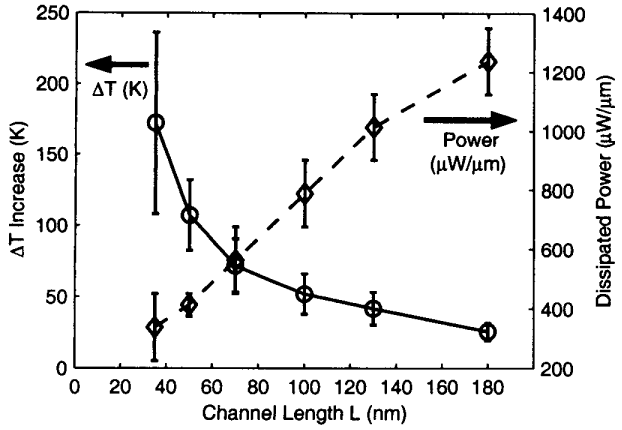

Fig. 8. Estimated average T rise above room temperature in the active area of the device under steady state operation (circles). The diamonds show the decrease in projected input power along the ITRS road map, but the temperature in the active region rises due to the increase and localization of power density.

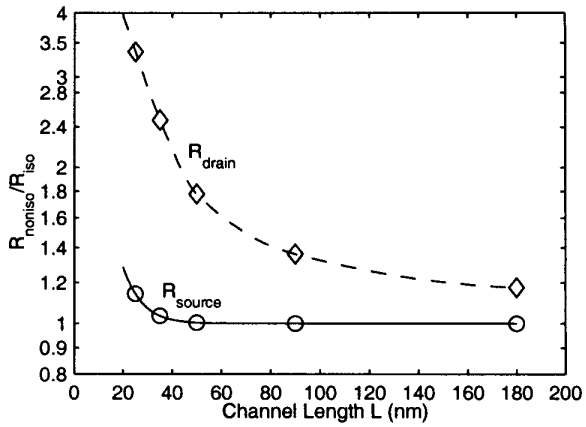

Fig. 9. Estimated hot spot effect on device source (circles) and drain (diamonds) series resistance normalized to their isothermal values (at room temperature). In the extreme limit of a 20 nm channel the drain series resistance may be magnified by a factor of 4 and the source resistance by a factor of 1.3 compared to their room temperature values.

continuum thermal issues and describe trends in their scalability. It is expected that increased electron scattering with hot phonons just inside the device drain will reduce effective electron temperatures, leading to less hot electron-induced oxide stress and degradation. The narrowed band gap near the drain however will increase impact ionization rates, creating increased substrate currents and device failures due to an earlier turn-on of the parasitic bipolar transistor. As device dimensions are reduced to scales almost comparable to the phonon hot spot, hot phonons emitted at the drain may affect the injection characteristics at the source (see Fig. 2). The source series resistance and the electron injection velocity are both functions of temperature [8] and would both have an impact on the maximum attainable drive currents. Fig. 9 is an estimate of the impact the localized hot spot may have on the source and drain series resistance along the ITRS road map. It is assumed that all hot spot phonons affect the drain resistance. To model any hot phonons which may travel to the source end, an exponential fall off function with decay length $\Lambda_{LO} = 10$ nm (on the order of the LO phonon relaxation length) was used. It can be extrapolated that in the extreme scaling limit, at channel lengths of 20 nm, the drain series resistance may be magnified by a factor of 4 and the source resistance by a factor of 1.3 compared to their isothermal (room temperature) values determined by the processing technology. An increase in R_{drain} is not commonly expected to impose a fundamental limit on CMOS scaling and performance [9], but any such effects on R_{source} and the electron injection velocity at the source end of the channel will impact the device current drive.

Conclusion

The generation and effects of drain-region hot phonons on CMOS device operation down to the ultimate scaling limits have been explored using BTE simulations of phonons based on a two-fluid model. It was argued that localized drain region phonon hot spots may affect the resistance and electron injection at the source end, hence the current drive of a quasi-ballistic device. Further work must be done to determine how the hot phonons emitted at the drain would affect the electron distribution, the injection velocity and the injection barrier on the source side of the channel. This work represents but a first attempt at quantifying such localized electro-thermal effects in ultra-scaled future technologies.

This work was done under SRC task 751.001. E. Pop is supported through the IBM/SRC Fellowship. The authors wish to thank S. Sinha, B. Winstead, M. Lundstrom, S. Laux, and K. Hess for helpful discussions.

References

[1] P. Sverdrup, Y.S. Ju and K.E. Goodson, "Sub-continuum simulations of heat conduction in silicon-on-insulator transistors", *J. Heat Transfer*, vol. 123, pp. 130-137, Feb. 2001

[2] Y.S. Ju and K.E. Goodson, "Phonon scattering in silicon thin films with thickness of order 100 nm", *Appl. Phys. Lett.*, vol. 74, pp. 3005-3007, May 1999

[3] G. Chen, "Nonlocal and nonequilibrium heat conduction in the vicinity of nanoparticles", *J. Heat Transfer*, vol. 118, pp. 539-545, Aug. 1996

[4] D. Long, "Scattering of conduction electrons by lattice vibrations in silicon", *Phys. Rev.*, vol. 120, pp. 2024-2032, Dec. 1960

[5] K. Hess, *Advanced Theory of Semiconductor Devices*, IEEE Press, 2000

[6] P. Sverdrup, S. Sinha, M. Asheghi, S. Uma and K.E. Goodson, "Measurement of ballistic phonon conduction near hotspots in silicon", *Appl. Phys. Lett.*, vol. 78, pp. 3331-3333, May 2001

[7] G. Dolling, "Lattice vibrations in crystals with the diamond structure", *Symposium on Inelastic Scattering of Neutrons in Solids and Liquids*, pp. 37-48, 1963

[8] M. Lundstrom, Z. Ren and S. Datta, "Essential physics of carrier transport in nanoscale MOSFETs", *Int. Conf. on Sim. Semic. Proc. and Dev. (SISPAD)*, Sep. 2000

[9] D.J. Frank, R.H. Dennard, E. Nowak, P.M. Solomon, Y. Taur and H.S.P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies", *Proc. IEEE*, vol. 89, pp. 259-288, Mar. 2001