Channel surfing

Nanoelectronics developers are aiming to boost channel properties in metal-oxide-semiconductor field-effect transistors (MOSFETs) as used in mainstream complementary MOSFET (CMOS = nMOS +pMOS) silicon technology. Dr Mike Cooke reports.

For many years CMOS gave both high performance and low cost via a single line of development, but the CMOS of today is not the CMOS of the 1960–80s (Figure 1). Pre-1990, modifications were mainly small variations on a well-established theme, making for much smoother development than has been seen in recent years. However, as development has become increasingly difficult, there has been a tendency to split out particular applications and optimize them.

The first major variation for higher-performance chips came in the 1990s, with aluminum being replaced by lower-resistance copper for the on-chip wiring (metal interconnects). A more extended and ongoing development has been lowering the dielectric constant (k) of inter-metal insulation to speed on-chip signal transmission. In the past couple of months, the companies Intel and IBM and research consortia such as Sematech have announced that the gate structure is ready for change to a high-k insulation and metal electrode, replacing the traditional polysilicon electrode and silicon dioxide insulator (which already has some nitrogen mixed in to increase the k value). And these are only the main lines of development.

Logic transistors can be optimized for high speed, low operating power or low standby power. The last two have come to the fore with the rise of battery-powered applications such as mobile phones. Memory producers have opted to focus on developing high density, often going beyond planar structures to three-dimensions, while continuing to use aluminum wiring rather than copper.

The implementation of new technologies is often delayed from the original speculation. The major reconstruction work needed results in workarounds and unexpected difficulties combining to push back the need or ability to use new structures and materials effectively. Further, a new technology needs a long-term future to justify implementation or even significant research funding.

Low-k dielectric materials is an example where far higher k values are being used compared with the predictions of 1999. The prediction of the 1999 International Technology Roadmap for Semiconductors (ITRS) for 2008 was for inter-metal k values of the order of 1.5. The 2006 ITRS update now expects effective k values in 2008 to lie in the range 2.7–3.0. There just aren't any materials that can be used to give a manufacturable k of 1.5 with reasonable yields at low cost. The traditional silicon dioxide inter-metal dielectric has a k value of 3.9.

A more recent example of delay has been the implementation of high-k dielectric with metal gate stacks.

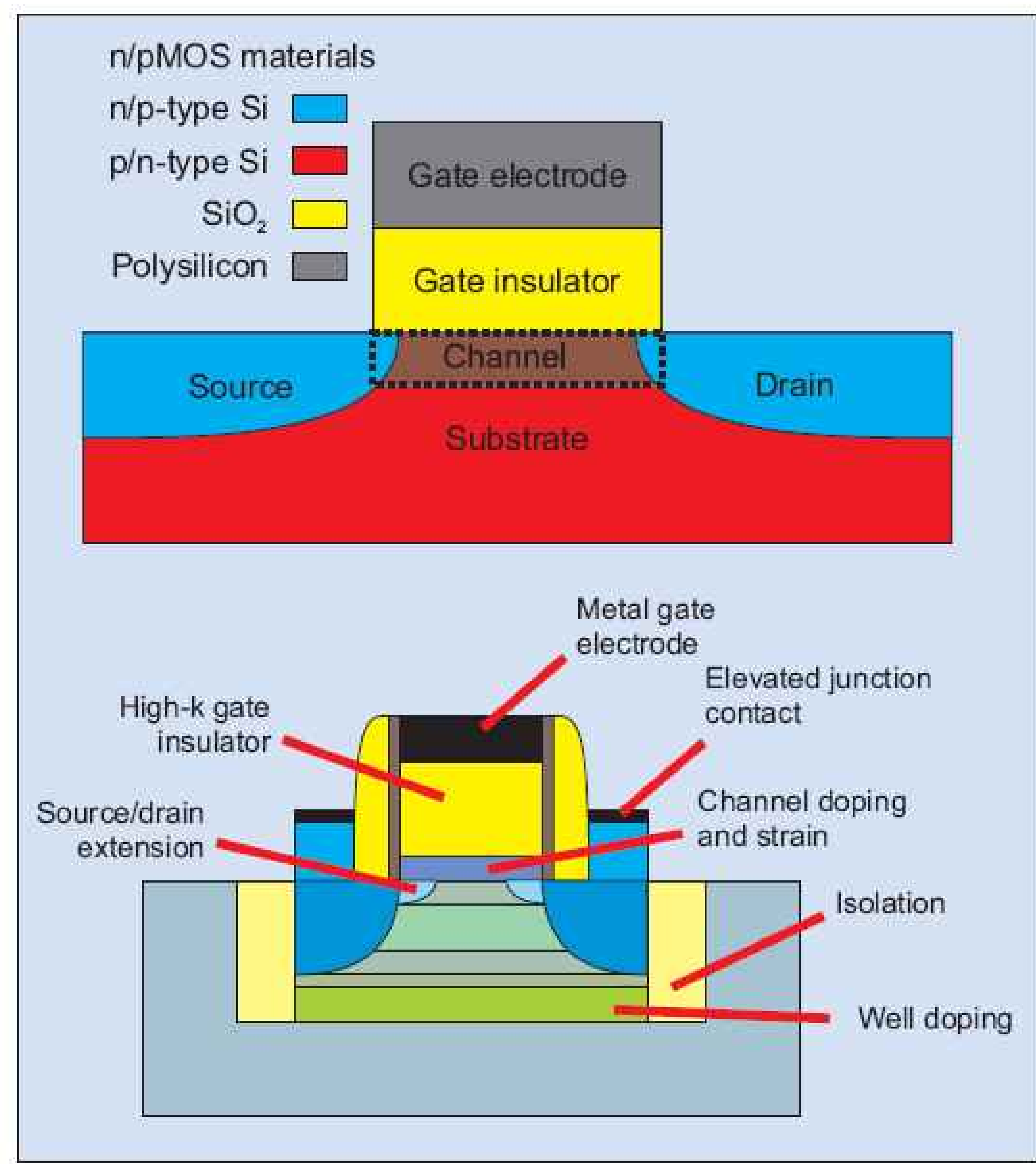


Figure 1. The CMOS of today is not the CMOS of the 1960–80s: traditional structure (top) and some of the new modifications (bottom).

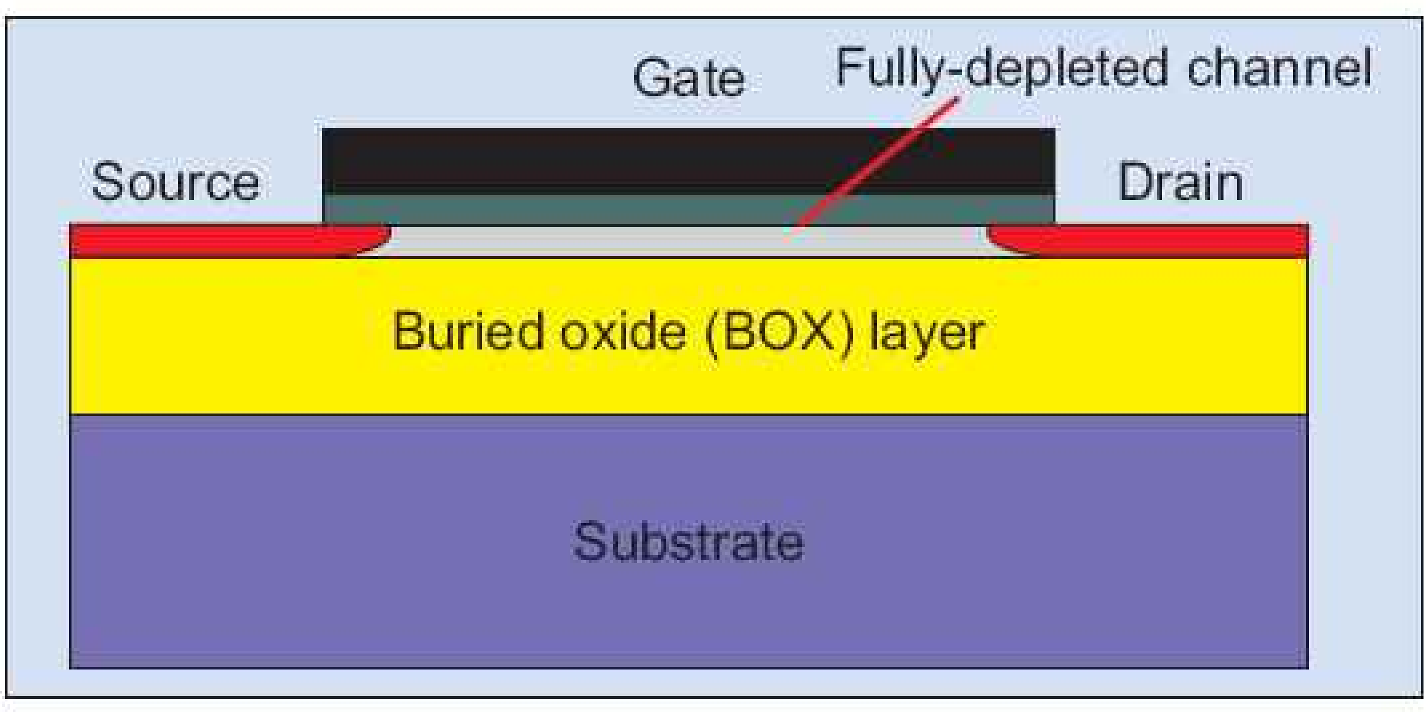


Figure 2. Ultra-thin body fully-depleted (UTB-FD) channel MOSFET.

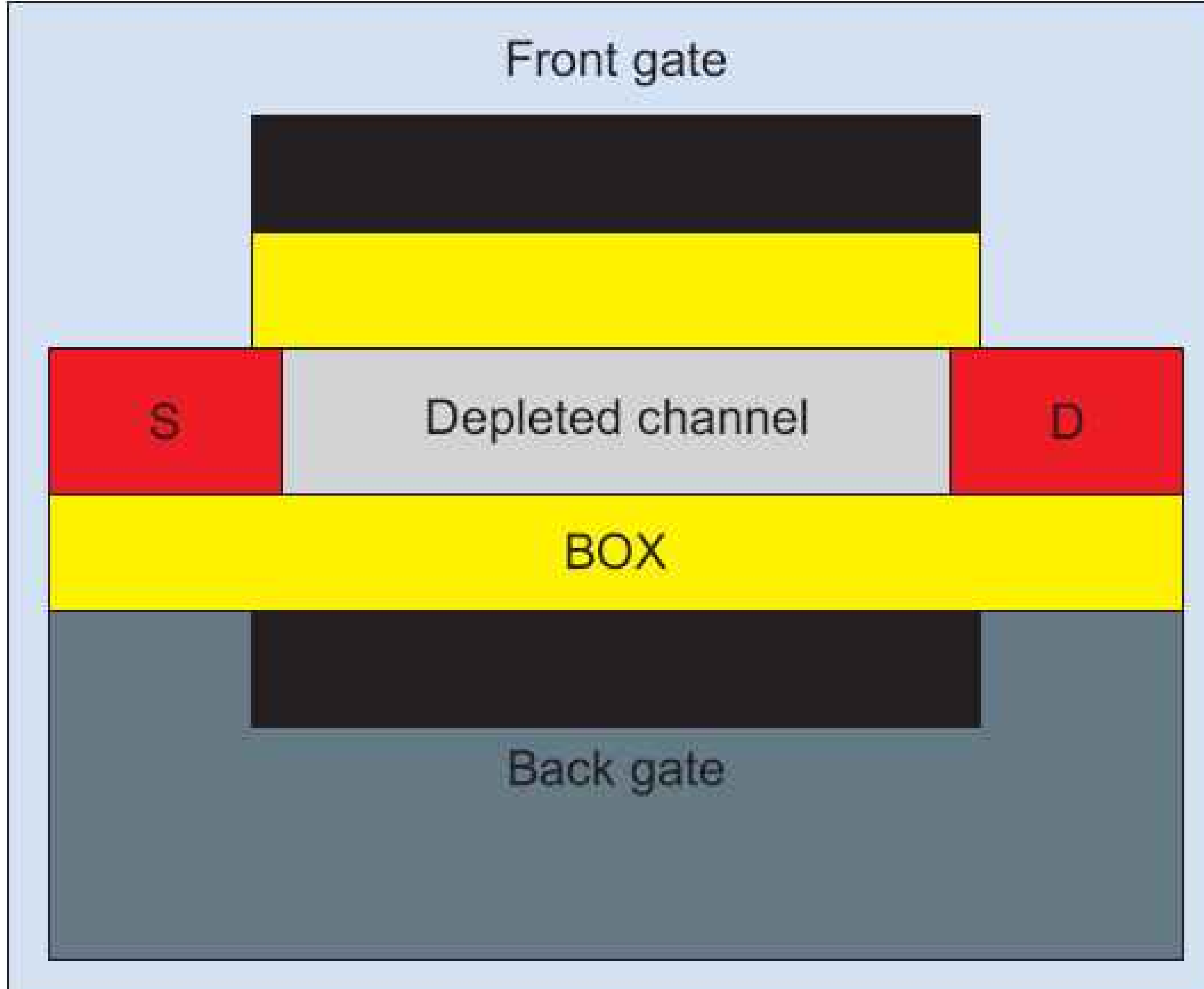


Figure 3. One dual-gate MOSFET structure.

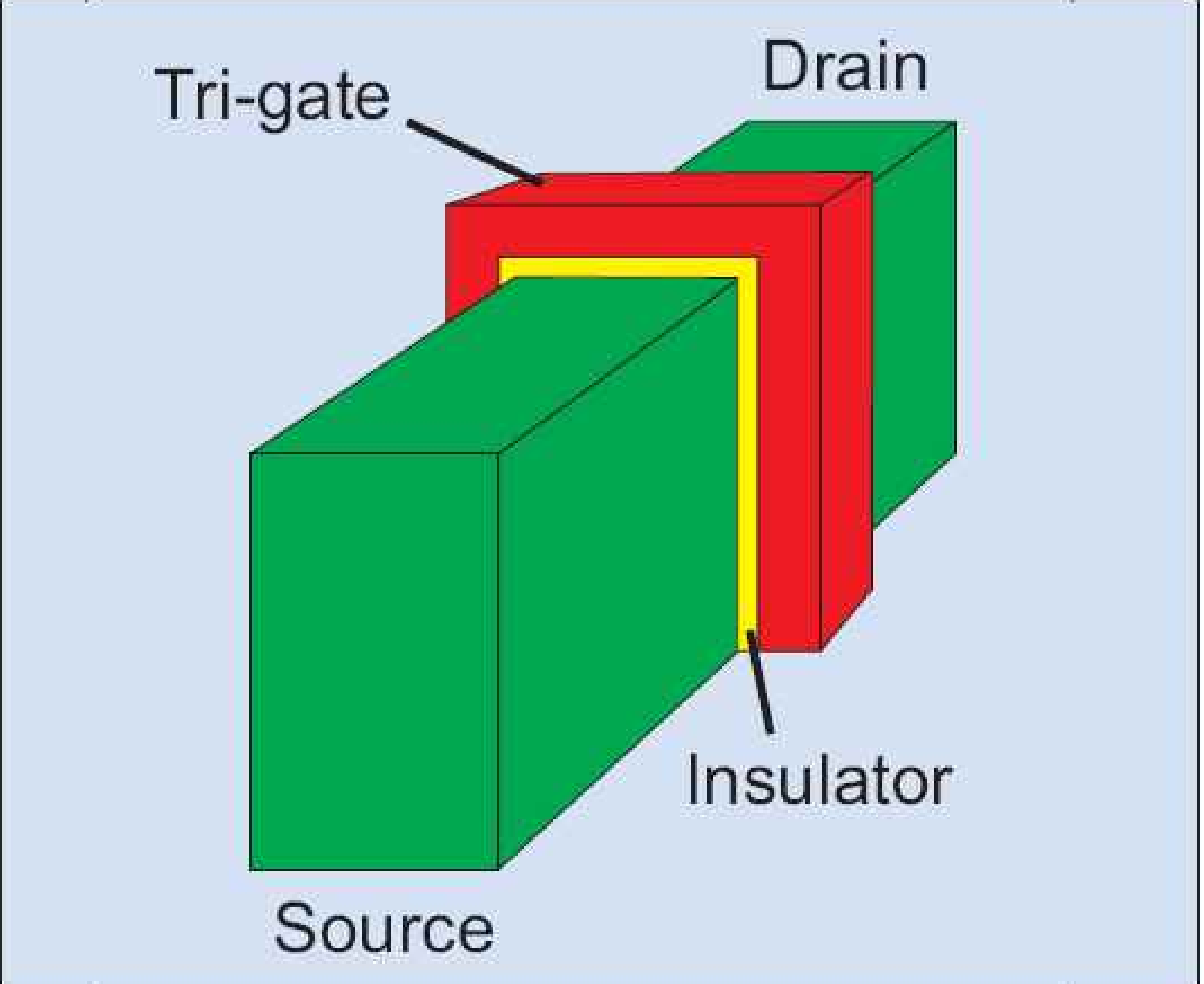


Figure 4. Tri-gate MOSFET as proposed by Intel, among others. In addition, tension (nMOS) or compression (pMOS) can be applied to the channel within the gate structure to increase mobility, along with metal/high-k gate materials.

The high-k with metal gate transistor is only now coming into production mainly for low-standby-power (LSTP) devices, while low-operating-power and high-performance implementations are now not due until 2010.

These delays have also knocked back other proposed changes, such as ultra-thin body fully depleted (UTB-FD) channel transistors (Figure 2). The 2006 update of the ITRS delayed the implementation of UTB-FD for high-performance logic from a 2008 schedule (given in ITRS 2005) to 2010. The low-power versions are due a couple of years later.

UTB-FD devices have a non-doped channel to reduce the number of ionized scattering centers. Not only does this give better mobility but, in short channels, it gives a quasi-ballistic enhancement, where some of the carriers travel from the source to the drain without any scattering events. The silicon layer is so thin that the channel is fully depleted of carriers until a gate potential is applied. However, the technology is limited in terms of how thin the silicon layer can be.

Beyond UTB-FD, dual-gate (DG) and more complicated structures are due to be used to shield and better isolate the source and drain regions. One DG structure consists of the usual (front) gate and a buried (back) gate to which a potential can also be applied (Figure 3). Other arrangements have vertical gates (side electrodes) or even gates that wrap around the channel — variously denoted tri-gate (Figure 4), FINFET or omega gate. The dual-gate channel is not doped, as with the UTB-FD structure, so DG structures are expected to benefit from ballistic enhancement too. DG MOSFETs are scheduled to scale out to 2020 and shorter channels (14nm node, physical gate length = 6nm), where the final ballistic enhancement factor is 2.11, while the expected mobility enhancement is only 1.03.

Another approach to dealing with short-channel effects (SCE), which lower the effective mobility of the channel due to high doping or electric field (saturation velocity) levels, is to change the channel composition to higher-mobility materials such as SiGe or a III-V compound semiconductor material (Table 1). Introducing strain can also massage mobilities upwards.

Along with implementations, there are also scheduled discontinuations of technology development: 'extended bulk planar' technology drops off the ITRS map in 2013 and UTB FD channel transistors in 2016. The DG transistor is seen as the 'ultimate MOSFET device', starting around 2011, with scaling projected out to 2020.

Table 1. Mobility and peak velocity comparisons for the conduction band [1].

Material	Mobility (10^3cm^2Vs)	High-field velocity ($10^7cm/s$)
Silicon	1	~1
Ge	2	~1
GaAs	5	1–2
$In_{0.53}Ga_{0.47}As$	10	~3.5
InAs	25	5 (?)
InP	3	3.5
GaSb	3	~1
InSb	77	4–5

Upward mobility

Traditional scaling rules for CMOS increases doping concentrations as device dimensions shrink. However, these rules are based on the simplistic view that

Gate
Source
High-k dielectric
Drain
$Al_{0.45}Ga_{0.55}As$ spacer
GaAs
$In_{0.25}Ga_{0.75}As$ channel
GaAs
$Al_{0.3}Ga_{0.7}As$ spacer
δ
Semi-insulating substrate

	Gate length		
	100.0	70.0	50.0
High-k dielectric	18.0	13.0	9.0
AlGaAs spacer	3.0	1.5	1.0
GaAs spacer	2.0	1.5	1.0
InGaAs channel	10.0	7.0	5.0
AlGaAs spacer	2.0	1.5	1.0
AlGaAs spacer	3.4	2.4	1.7
δ-doping	2.65	3.50	3.25

Figure 5. Implant-free III-V MOSFET simulated by Kalna et al [3]. The table shows layer thicknesses (in nm) and the δ-doping concentration (in $10^{12}cm^{-2}$) for various gate lengths (in nm).

increased doping leads to increased conductivity because mobility is approximately constant, i.e. that the conductivity is proportional to the number of free carriers released by ionized impurities. Unfortunately, among the effects of high doping concentration is the increase in scattering from ionized impurities, reducing mobility severely. Between impurity concentrations of $10^{16}cm^{-3}$ and $10^{20}cm^{-3}$, mobility in silicon falls by almost three orders of magnitude. Another problem with increased doping is that the impurities begin to clump together, creating non-uniformities. Further the electric field in short channels more easily enters the saturation region, where increased fields do not lead to increased carrier velocities. Although the scaling rules look to reduce operating voltages so that the electric field remains below saturation, this is often not possible in short-channel structures, where two-dimensional corrections to one-dimensional models become important (e.g. the field increases near sharp structures).

Mobility enhancement using strained structures of silicon and silicon germanium have already been implemented, first for high-performance transistors around 2004, where higher fields are used to boost performance, and hence carrier velocities may enter the region where they cannot go any higher (i.e. they saturate). To increase electron mobility, a thin layer of silicon is deposited on SiGe, which has a larger lattice constant due to the larger germanium atoms, and hence is in tension. For hole mobility enhancement, a strained layer of SiGe on bulk silicon can be used to provide compression.

However, as scaling continues, the modest improvements in channel mobility provided by strain will not be sufficient and it is expected that new materials such as germanium or III-V compound semiconductors — and, beyond that, 'nanowires' or even carbon nanotubes — will be needed.

Last year, the US Semiconductor Research Corporation (SRC) — an industry-funded university-research consortium for semiconductors and related technologies — launched a Non-Classical CMOS Research Center in order to research III-V materials to extend CMOS capabilities. A team of five universities is to collaborate for three years with more than $7m of funding. Depending on the results, two further years of funding may be agreed.

The aim is to enhance CMOS gate speeds with lower power dissipation and to beat the ITRS schedule for alternative channel materials, with significant impacts on chip manufacturing expected as early as 2012–2014 (32nm). The ITRS plan is for alternative materials to be available for semiconductor production at the 22nm level (2016–2019).

The Non-Classical CMOS Research Center is being led by University of California, Santa Barbara. The four other institutions are Stanford University, University of California, San Diego, the University of Massachusetts-Amherst, and the University of Minnesota.

Applications that should benefit could include communications, computing, gaming, automotive and consumer electronics. The research is partly inspired by the technology and footprint improvements provided by compound semiconductors in communication devices such as satellite dishes, where indium gallium arsenide-based preamps have shrunk dish diameters from 1.5m to 0.5m in less than 10 years, while doubling reception quality.

Unfortunately, CMOS devices are delicately structured — change the channel material and there are implications for the gate stack and other features. For example, the silicon dioxide, or even the new hafnium dioxide (HfO_2) high-k dielectric gate insulators now on silicon CMOS pilot lines, would probably have

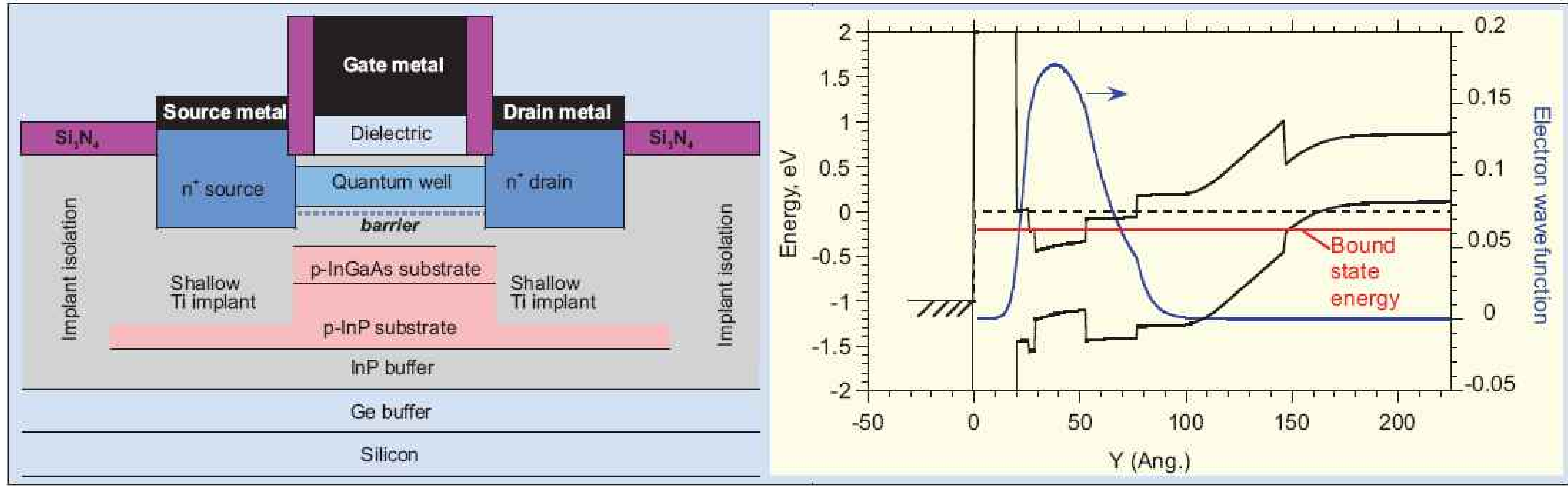


Figure 6. Lattice mismatches between III-V materials and Si create the need for buffer layers: (left) a possible III-V nMOSFET on silicon [4]; (right) graph of quantum well confinement scheme in the vertical direction.

to be changed, since surface states easily arise at the dielectric/channel interface. Indeed, the rise of silicon CMOS was due in no small part to the ease with which the Si/SiO_2 interface could be made to avoid or mitigate these problems. III-V FETs usually contact a metal gate directly to the semiconductor (metal–semiconductor FET, or MESFET).

Although this lack of a native oxide for III-V materials has been a barrier to the successful development of III-V MOSFETs for 30 years, Freescale Semiconductor's Matthias Passlack recently reported progress in assessing $Gd_xGa_{0.4-x}O_{0.6}/Ga_2O_3$ dielectric stacks on GaAs [2]. The relative dielectric constant is 20.8±1 and the breakdown field exceeds 4MV/cm. At an electric field of 1MV/cm, the leakage current density ~$2x10^{-8}A/cm^2$. The interface state density (D_{it}) has a broad, u-shaped minimum of less than $2x10^{11}cm^{-2}eV^{-1}$ on n-type GaAs.

Last year, Passlack joined with Kalna and Asenov of the University of Glasgow in the UK to create a Monte Carlo simulation of an implant-free InGaAs MOSFET (Figure 5). The reason for developing an implant-free MOSFET is that earlier simulations of structures with implants showed that, while 80nm $In_{0.2}Ga_{0.8}As$ MOSFETs would easily outperform Si and strained Si devices, the marginal improvement at 35nm drops dramatically. The implant-free $In_{0.25}Ga_{0.75}As$ MOSFET promises better behavior with vertical and lateral scaling. Moving from a 100nm to a 70nm device delivers a 60% drain current increase and a maximum transconductance of 2080mS/mm. Moving to 50nm, the drain current is 90–100% better than the 100nm version, with a maximum transconductance of 3190mS/mm.

For processing, the standard silicon CMOS infrastructure does not yet exist in an equivalent III-V form tailored to ultrahigh-volume manufacturing rather than niche applications. In addition, semiconductor manufacturers like Intel are fairly insistent that any channel change has to integrate onto a silicon substrate, which is not an easy problem, given the large lattice mismatch in general (Figure 6). Intel has worked with Qinetiq in the UK on indium antimonide (InSb)

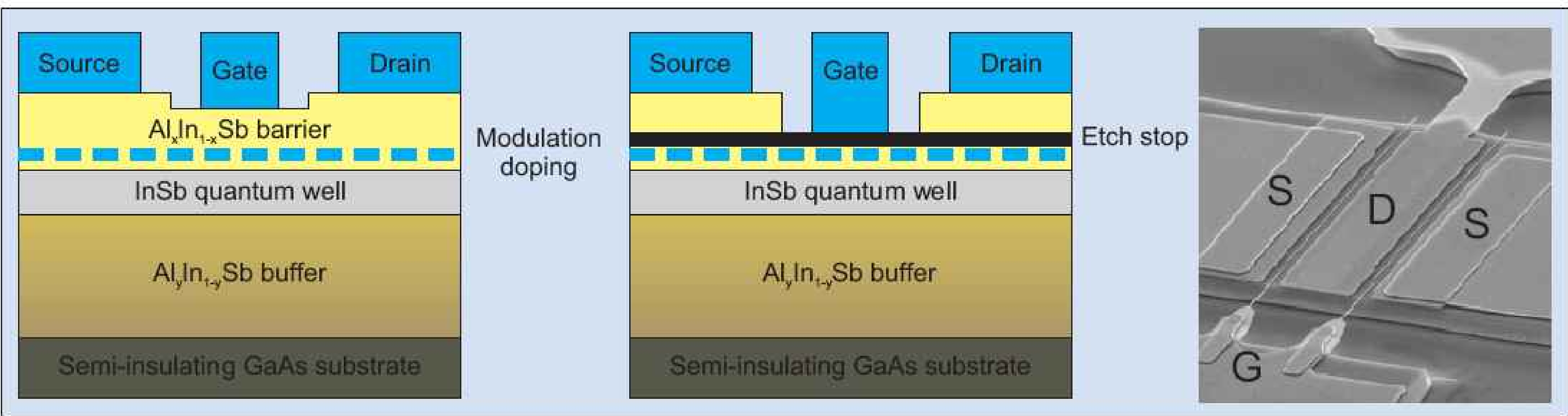


Figure 7. InSb quantum well transistor developed by Intel and Qinetiq [5]. For the 85nm gate length device reported in 2005, the layers were a 3µm AlInSb buffer, a 20nm InSb quantum well, a 5nm thick $Al_{0.2}In_{0.8}Sb$ spacer, δ-doping, and a 45nm $Al_{0.2}In_{0.8}Sb$ barrier with an optional etch stop. The figure shows devices that are (left) depletion-mode (normally-on) and (middle) enhancement-mode (normally-off). The δ-doping consists of a Te donor layer (density $1–1.8x10^{12}cm^{-2}$). The researchers were able to achieve unity gain cut-offs (f_T) of 305GHz (enhancement) and 256GHz (depletion) with V_{DS} = 0.5V. This is a 50% boost over equivalent silicon nMOS devices while consuming 10x less active power.

quantum well transistors (Figure 7). However, it has not yet reported the realization of its dream of putting such structures onto silicon.

Other drawbacks from the physics of III-V materials include densities of states that are lower than for silicon, with adverse implications for drive current and parasitic capacitances. Lower carrier effective masses in III-Vs can also lead to poor confinement for sub-22nm devices, band–band tunneling, and off-state leakage effects.

Beyond the beyond

III-V or even II-VI materials (such as zinc oxide, ZnO) are still likely to be in the frame for smaller dimensions when nanowire channels are developed. However, it is by no means certain that this is the way the industry will go. Another possibility is carbon nanotubes, where the chicken-wire structure of graphite (due to sp^2 bonding, instead of diamond sp^3 bonding) is rolled into tubes. However, so far, due to the difficulty of working with such small objects, R&D nanotube devices have depended on chance alignments with electrode structures.

III-V or even II-VI materials (such as zinc oxide, ZnO) are still likely to be in the frame for smaller dimensions when nanowire channels are developed.

In the past few years, atomically thin layers of graphitic film, or graphene, have been grown on SiO_2/Si and other substrates, with researchers at the University of Manchester in the UK in the vanguard of this research. Two of the researchers, professor Andre Geim and Dr Kostya Novoselov, give details of a graphene transistor in a review article on progress in graphene studies [6]. The transistor structure is only one atom thick and less than 50 atoms wide. The scientists believe that this could allow the rapid miniaturization of electronics to continue when CMOS technology runs out of steam.

Graphene structures have been grown on top of non-crystalline structures, in liquid suspension, or as suspended membranes. Epitaxial growth on silicon carbide (SiC) produces high-mobility charge carriers, which is important for electronics applications.

Geim does not expect that graphene-based circuits will come of age before 2025. Until then, silicon technology should remain dominant. But he believes graphene is probably the only viable approach after the silicon era comes to an end: "This material combines many enticing features from other technologies that have been considered as alternatives to the silicon-based technology. Graphene combines the most exciting features from carbon-nanotube, single-electron and molecular electronics, all in one."

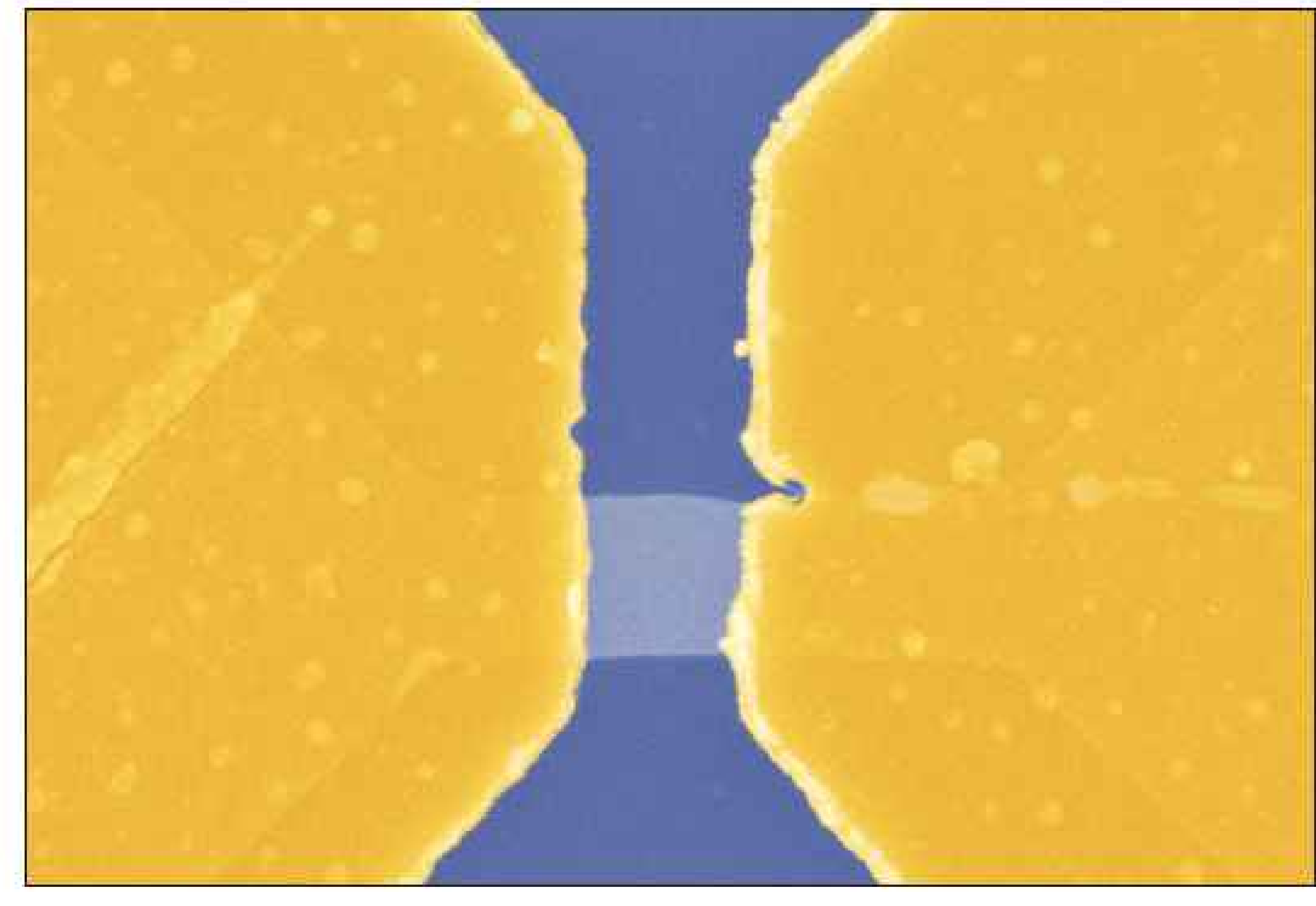
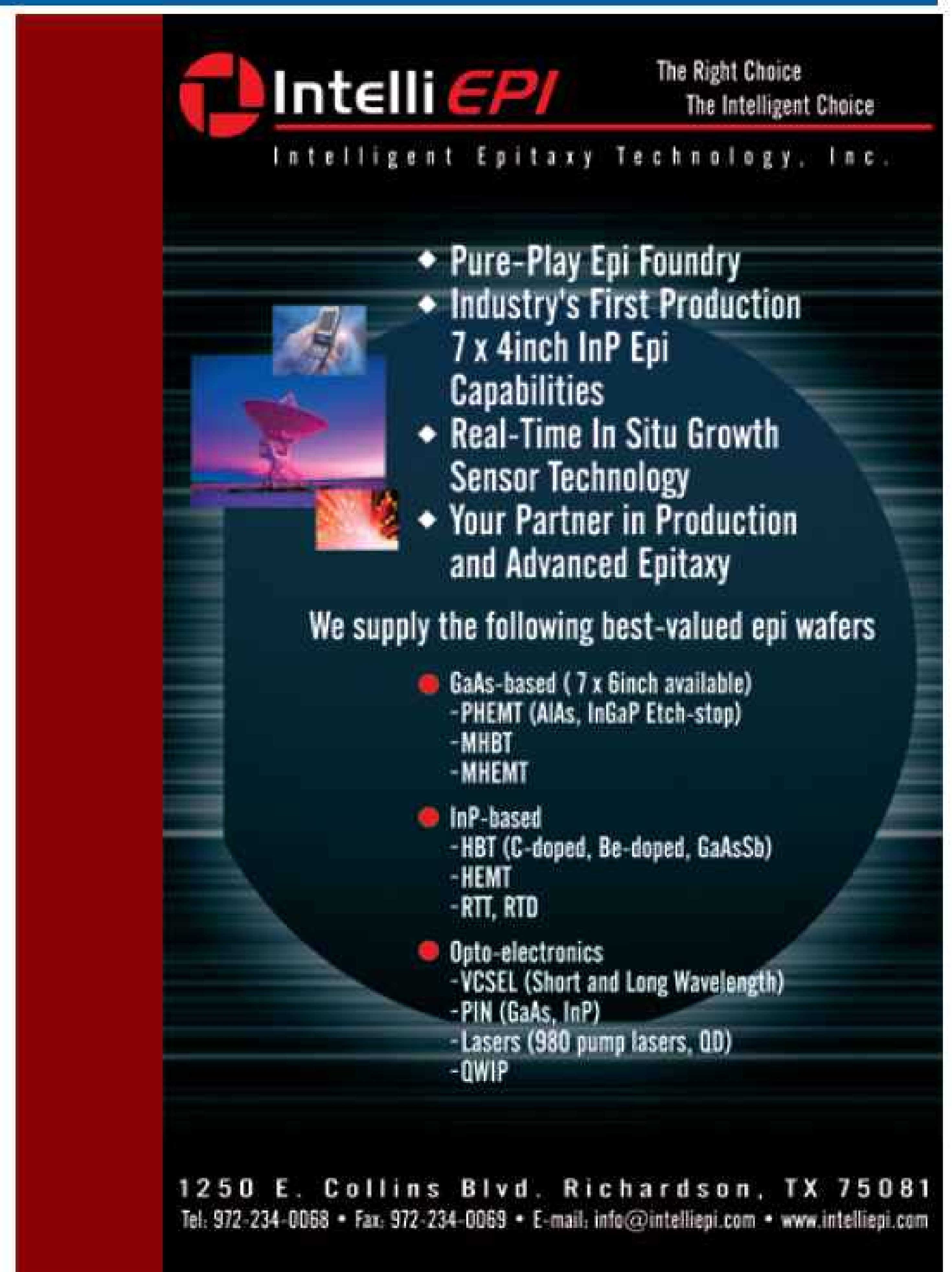


Figure 8. Scanning electron micrograph (in false color) of a transistor made on single-layer graphene. Au contacts shown in gold; SiO_2/Si substrate in blue. The width of the graphene wire is 200nm.

The transistors are made by carving nanoribbons, semi-transparent conduction barriers and quantum dots all in the same graphene layer (Figure 8). The conductance can be controlled by gates (back or side electrodes, also constructed from graphene).

"We have made ribbons only a few nanometers wide and cannot rule out the possibility of confining graphene even further — down to maybe a single ring



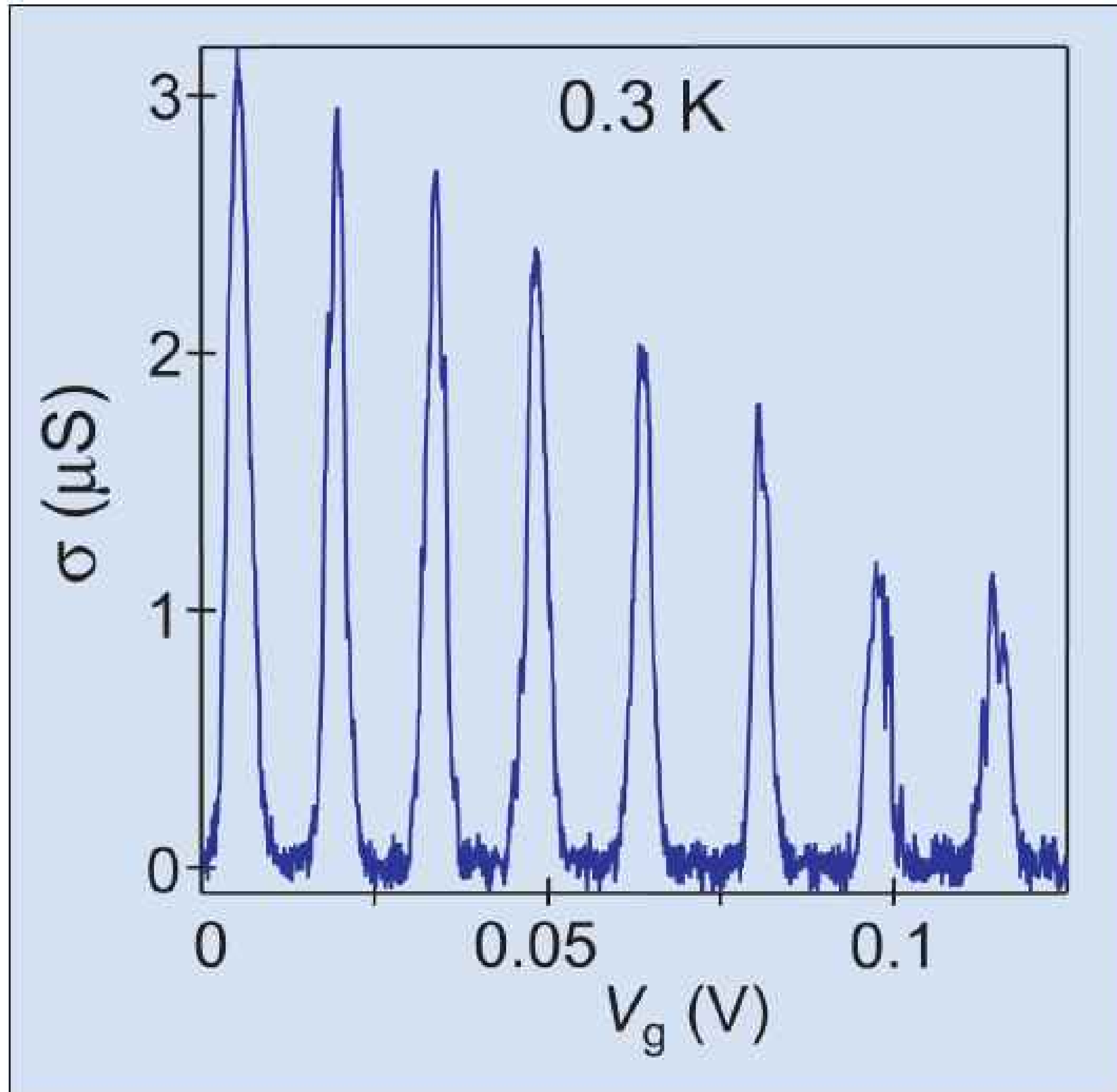


Figure 9. Coulomb blockade of conductance at low temperature (0.3K).

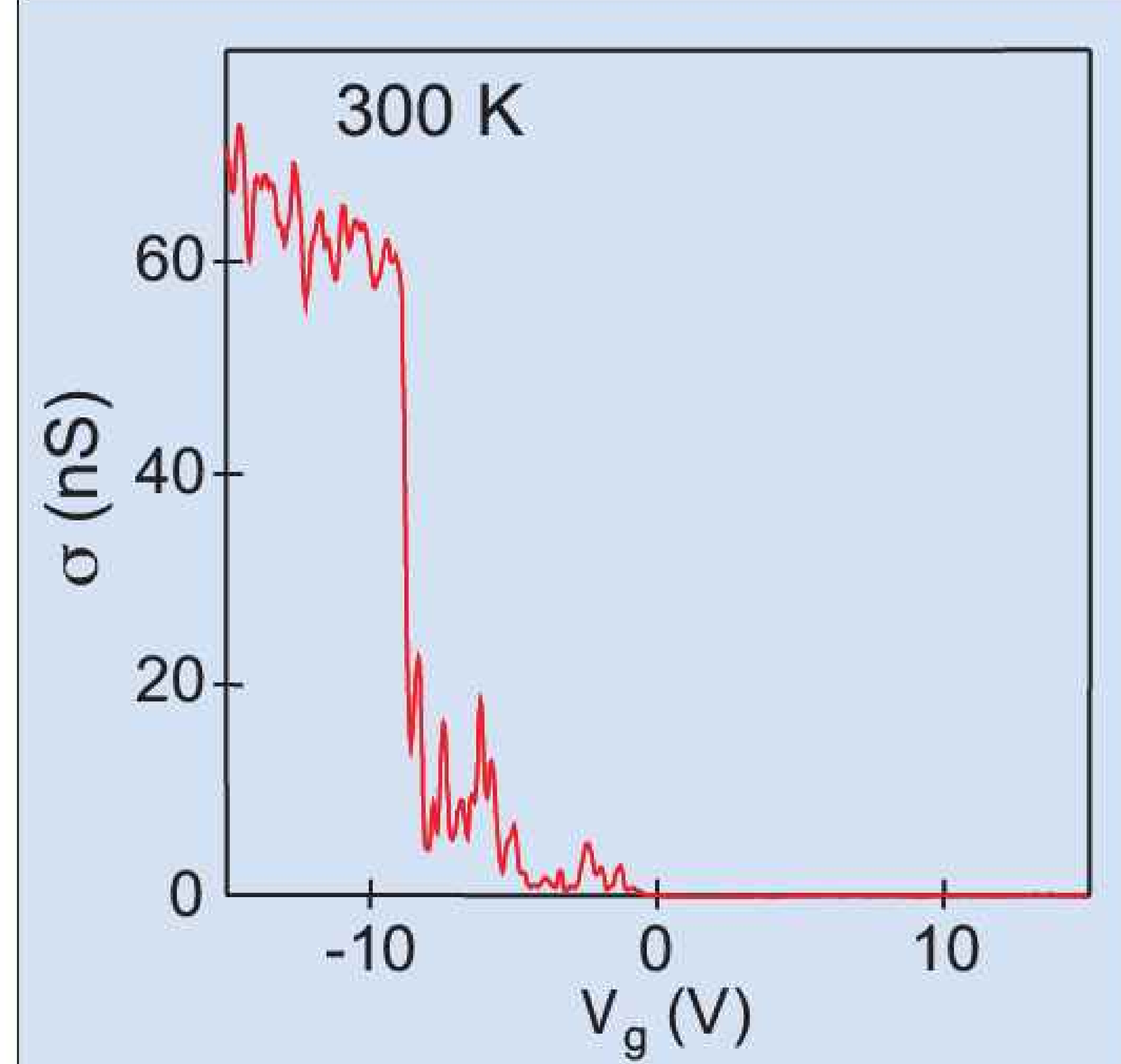


Figure 10. Conductance pinch-off at room temperature.

of carbon atoms," says Geim.

Two graphs of the transistor action for two structures are presented in their article published in the journal Nature Materials. One shows coulomb blockade effects (demonstrating the effects of single charge carriers) in a relatively large quantum dot (0.25μm in diameter) at low temperature (Figure 9). Another graph shows the operation of a device consisting of 10nm-scale structures at room temperature, showing complete pinch-off of the current (Figure 10). The caption for the figure promises that a more detailed presentation is currently "in preparation".

"At the present time, no technology can cut individual elements with nanometer precision. We have to rely on chance by narrowing our ribbons to a few nanometers in width," says Dr Leonid Ponomarenko, who is leading this research at Manchester. "Some of them were too wide and did not work properly, whereas others were over-cut and broken."

But Ponomarenko is optimistic that this proof-of-concept technique can be scaled up: "To make transistors at the true nanometer scale is exactly the same challenge that modern silicon-based technology is facing now. The technology has managed to progress steadily from millimeter-sized transistors to current microprocessors with individual elements down to tens of nanometers in size. The next logical step is true nanometer-sized circuits, and this is where graphene can come into play because it remains stable — unlike silicon or other materials — even at these dimensions."

Geim and his colleagues discovered graphene about two years ago. The Manchester group discovered graphene by using a non-epitaxial process on a SiO_2/Si substrate.

"Our approach is not scalable but it is good enough for research and proof-of-concept devices," says Geim. Other groups use an epitaxial process on silicon carbide that could lead the way to large-scale industrial application, although the present quality of the resulting graphene layers is not good enough for this.

The first graphene-based transistor was reported by the Manchester team at the same time as the discovery of graphene [7], and other groups have reproduced the result (for example, the USA's Georgia Institute of Technlogy has grown the graphene on SiC substrates, [8]). However, these graphene transistors were very 'leaky'.

The new transistors are much less leaky and Geim et al have also shown that graphene remains highly stable and conductive, even when it is cut into strips only a few nanometers wide. All other known materials — including silicon — oxidize, decompose and become unstable at sizes tens times larger. The poor stability of these other materials has been a fundamental barrier to their use in future electronic devices, limiting microelectronics development. ■

References

[1] From Ioffe Institute web-site, www.ioffe.rssi.ru/SVA/NSM/Semicond/
[2] Passlack, J. Vac. Sci. Tech., vol. B23, p. 1773, 2005.
[3] Kalna et al, J. Phys.: Conf. Series vol. 38, p. 200, 2006.
[4] From presentation at UCSB Engineering Insights Meeting, October 2006, by Professor Mark Rodwell.
[5] Datta et al, International Electron Devices Meeting (IEDM) 2005, session 32.1.
[6] Geim and Novoselov, Nature Materials, p.183, March 2007.
[7] Novoselov et al, Science, vol. 306, p. 666, 22 October 2004.
[8] Berger et al, J. Phys. Chem. B, vol. 108, p. 19912, 2004.