Flash fast forward to quantum dot memory

Non-volatile memory in the form of NAND Flash is now driving development in the silicon semiconductor industry. However, this technology can only continue on the back of adaptation to ever higher densities. Dr Mike Cooke surveys the coming changes and looks at how III-V quantum dots offer the prospect of fast non-volatile memory.

A bewildering array of technologies has been used to support the memory and storage needs of electronics over the years, from paper tapes and punched cards to semiconductor-based products. Many factors play a part in deciding which technology is used for a particular level of storage; some of the most important are read/write speeds, storage density, endurance, reliability and, finally and often most importantly, cost. These requirements, like so much in life, are usually contradictory. The engineering problem is to find the optimal combinations for specific applications.

In a PC, the memory close to the central processor unit (1st and 2nd level cache) must be fast but not particularly dense or high capacity. The technology of choice for these applications is static random access memory (SRAM), which is volatile (data disappear when the equipment is turned off). The working memory of a PC needs high density and high capacity in addition to high speed. Dynamic random access memory (DRAM) meets these requirements, but is even more volatile than SRAM — it needs to be constantly reminded of the information it contains since its retention is of the order of milliseconds.

To retain information when equipment is turned off requires some form of nonvolatile memory. In the PC, the standard technology for this is the magnetic hard disk drive. This provides high capacity, a moderate density of information, but a slower read time. More recently, semiconductor-based Flash memory has become increasingly popular for highly compact consumer applications such as mobile phones and other portable devices such as mp3 players and digital still cameras. Flash offers high and increasing storage densities, but rather slow write and erase times. This is a disadvantage in many applications such as video camcorders. Although Flash is used for such video storage, the device usually contains a DRAM buffer to contain the raw data, which then takes some time to transfer to nonvolatile form when filming stops, often blocking continued use of the camera until the process is finished.

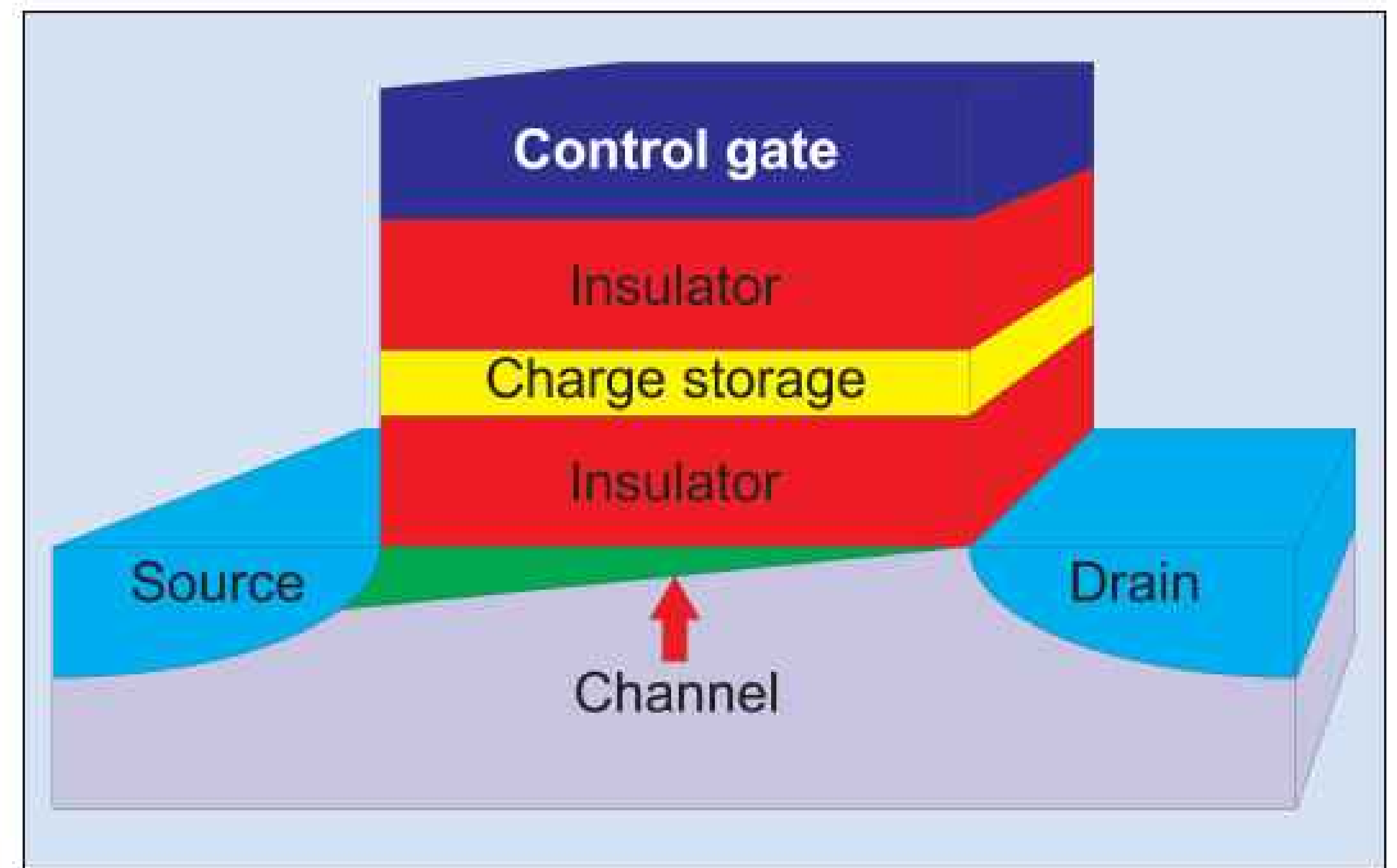

Figure 1. Schematic of Flash memory transistor.

Flash storage

A basic Flash memory transistor is similar in structure to those used in complementary metal oxide semiconductors (CMOS) (Figure 1). The main addition is a charge-storage layer between the gate electrode and the channel. The presence or otherwise of charge in this layer affects the electric field in the channel and hence the threshold voltage needed on the control gate for increasing the channel's carrier concentration and hence its conductivity. The read operation for the memory/charge state consists effectively of a relatively simple and fast current-sensing measurement. If a number of different charge states can be separated with different threshold-voltage windows, multi-bit memory devices become possible.

The difficult part for operating Flash is to change the memory state — that is, to write or erase charge from the storage layer. This operation stresses the material structure and leads to degradation and failure over time. In general, the transfer of holes (in contrast to electrons) is more damaging, so commercial Flash memory and related devices depend on electron storage. One needs to arrange the structure so that it can retain charge for long periods (more than ten years) and can be rewritten many times.

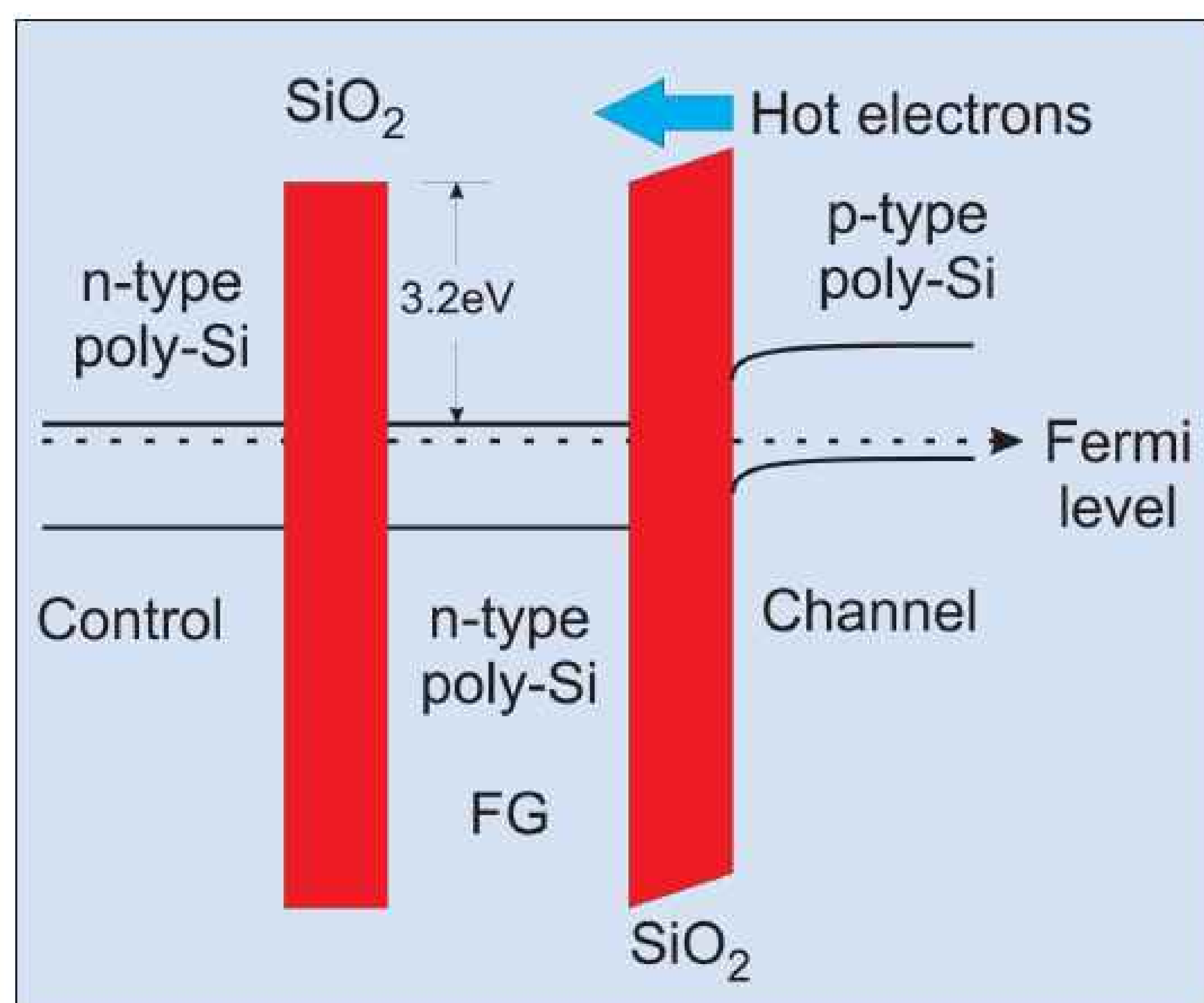

Figure 2. Band structure of floating-gate (FG) Flash memory. Hot electrons are those with enough energy to cross the insulating barrier.

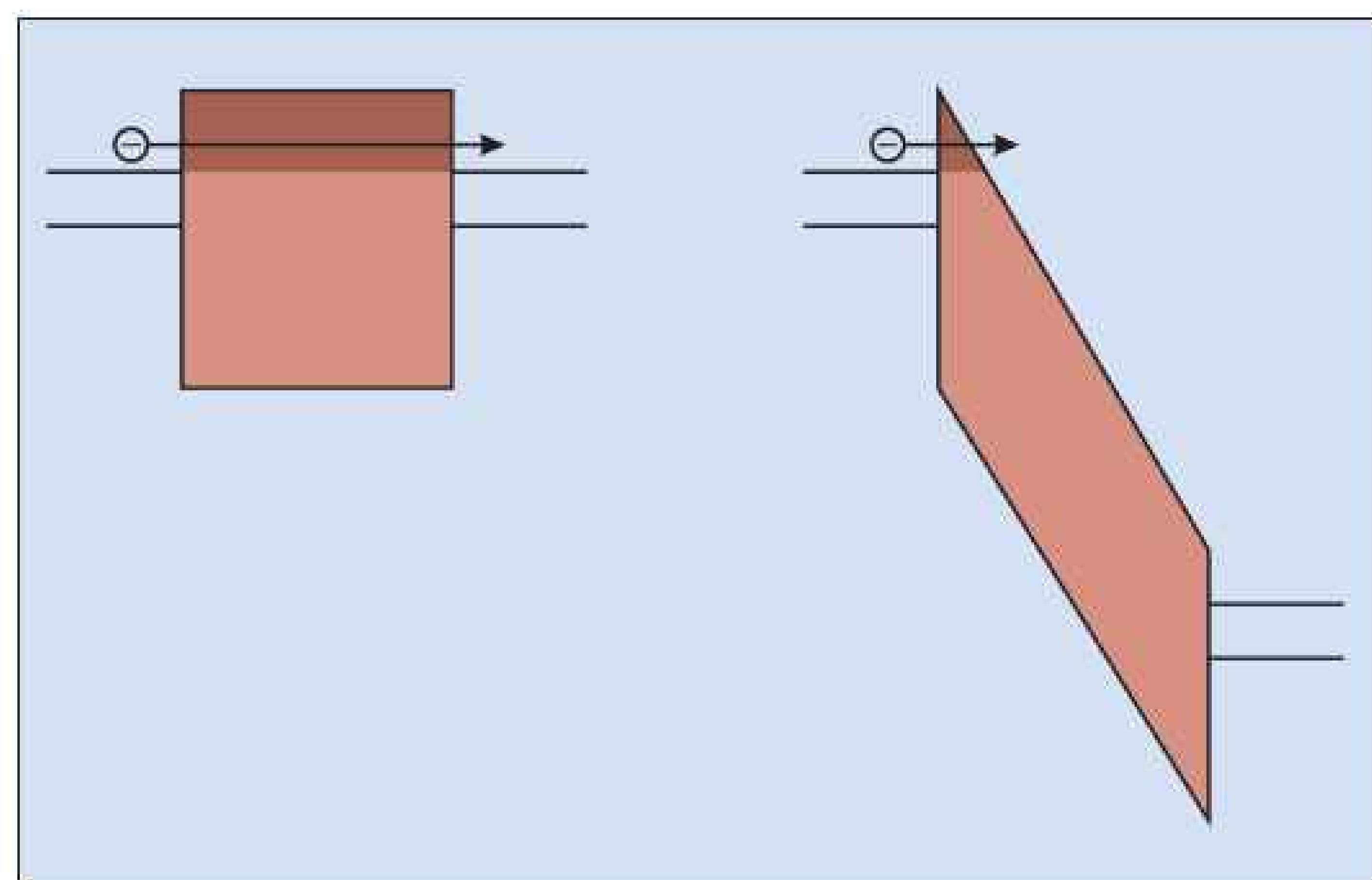

Figure 3. Fowler–Nordheim tunnelling (right): applying an electrical field across an insulating barrier reduces its effective thickness, allowing the transmission of charge out of the storage layer.

Faster write/erase operations lead to increased stress and hence greater device degradation. Present Flash devices are limited to write/erase operations of at best tenths of milliseconds, while the systems that they service often have clock speeds of up to gigahertz (GHz, or 1000MHz). Of course, things can be sped up by arranging parallel write/erase operations. Hence, one can have devices with 20Mbit/s and even 80Mbit/s data transfer rates. The different arrangements come under such labels as NOR and NAND Flash, which have different strengths and weaknesses from an applications perspective. NOR offers true random access of memory bits, but lower numbers of write/erase cycles, and is often used to hold program code that needs upgrade options (firmware). Denser NAND is preferred for holding data such as digital images, music and video.

A traditional Flash memory storage layer consists of a conducting polysilicon 'floating gate' that is insulated from the control gate electrode and the conducting channel by an insulator such as silicon dioxide dielectric. A variation is SONOS memory, which uses a stack consisting of silicon-oxide-nitride-oxide-silicon where charge is stored on trap states in an insulating silicon nitride (or oxynitride) layer. In floating-gate Flash, ONO (oxide–nitride–oxide) is also often used instead of the SiO_2 insulator, but then one must be careful that charge storage is on the gate and not in the nitride layers.

Two main methods are used to get electrons in and out of the charge storage layer: hot-carrier injection and tunneling. 'Hot carriers' are those in high enough energy states to cross the insulating barriers into the floating gate or nitride charge storage regions. Silicon dioxide, for example, offers an energy barrier of about 3.2eV to electrons between the electrode and floating gate (Figure 2).

One way to create hot carriers is to accelerate electrons from the source to drain and arrange for a control gate potential to deflect them onto the charge storage region. 'Tunneling' refers to the quantum-mechanical effect where part of the electron wave-function can trespass into classically forbidden regions (such as insulation barriers) and hence allows some transmission of electrons — the thinner the barrier, the greater the transmission. For charge retention, one needs the barrier to be sufficient to block off such transmission but, for write/erase operations, applying an electrical field reduces the barrier's effective thickness, allowing charge to escape. This technique is referred to as Fowler–Nordheim (FN) tunneling (Figure 3).

Since it is difficult to use hot-carrier injection to erase the charge state, devices tend to use tunneling for this operation. Tunneling can also be used to push charge into storage, but can be stressful for the device's performance. NAND Flash tends to use tunneling for both write and erase operations, while NOR will often use hot-electron injection from the channel (channel hot-electron injection, or CHEI) to write charge onto the charge trapping layer. System operating voltages of 1.8–5V are converted to the higher voltages of 15–17V that are needed for write/erase by 'charge pump' circuitry.

When first introduced in the late 1980s, Flash tended to lag behind DRAM in implementing advanced technology and consequent density increases. However, in recent years the situation has reversed, with NAND Flash now leading the way. The International Technology Roadmap for Semiconductors (ITRS 2007) puts the expected density for DRAM 'at production' this year at 3Gbits/cm^2, while in 2015 it is expected to reach 15Gbits/cm^2. Flash memory densities in 2008 are expected to be of the order of 8Gbits/cm^2 for single-level memory cell devices and 17Gbits/cm^2 for two-level cells. The 2015 figures are 40Gbits/cm^2 and 90Gbits/cm^2, respectively.

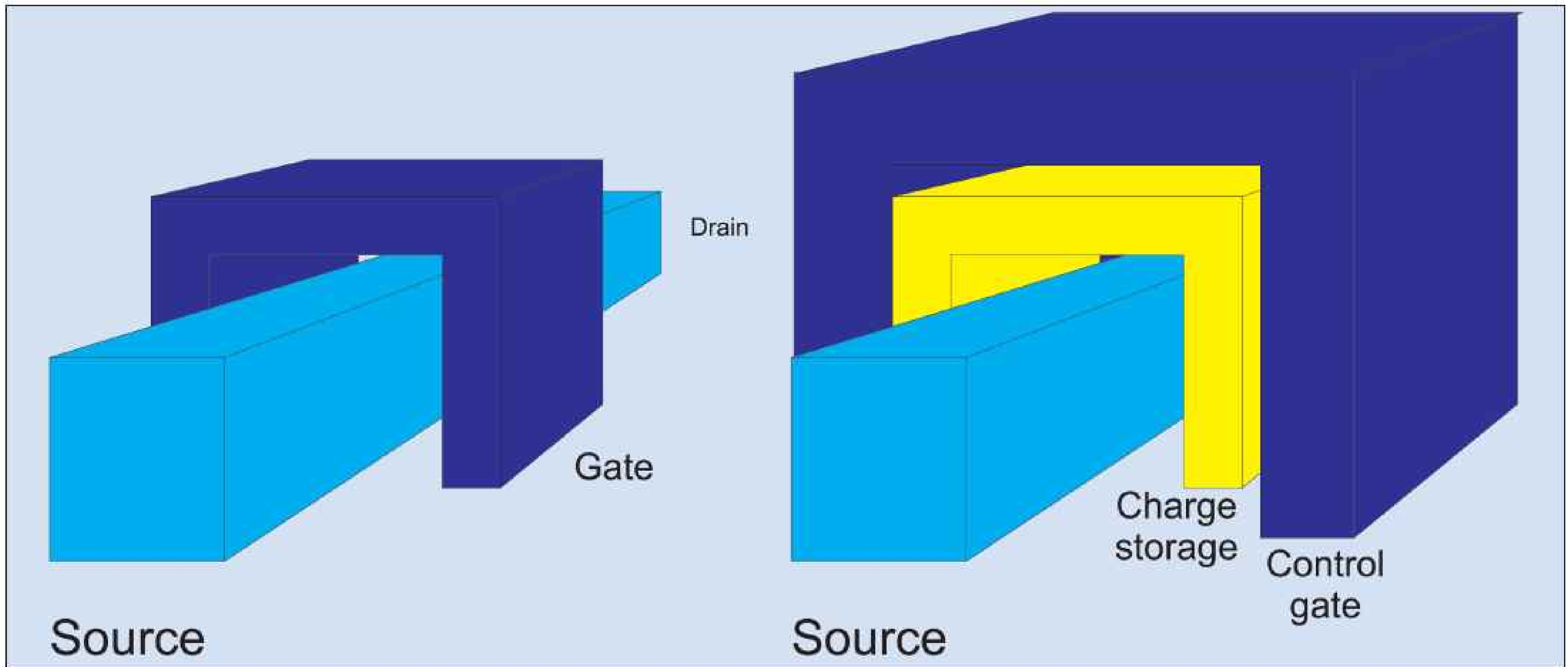

Figure 4. Schematic diagrams of FinFET (left) and FinFlash (right).

A couple of years ago the future was looking bleak for traditional Flash memory, and a number of nonvolatile alternatives were being proposed. The latest International Technology Roadmap for Semiconductors [1] reveals that many of the near-term roadblocks have in theory been cleared.

This has been achieved largely through slight variations of Flash. For example, floating-gate NAND Flash is expected to be pushed aside as the majority technology in favor of some form of charge trapping in a silicon nitride (SiN) layer in about 2010 (the 45nm technology node). Floating-gate NAND Flash could continue to find makers, so long as the ONO insulator structure presently used can be replaced by a high-k dielectric, as is currently taking place in the high-performance CMOS logic domain.

For charge trapping, the two dielectric layers are expected to consist of either traditional SiO_2 or ONO (tunnel) and Al_2O_3 (blocking). The control gates may move to metal from polysilicon, resulting in a metal–Al_2O_3–nitride–oxide–silicon (MANOS) structure. Continued density increase is expected from two-level devices in about 2013, in addition to each transistor being able to store four bits (16 states) from about 2010. The current number of bits per cell is two, with three-bit/cell memories expected next year. All these devices are aiming at nonvolatile data retention of 10–20 years and endurance of 10^5 write/erase cycles.

Although the ITRS and the manufacturing companies seem to have a workable plan for developing Flash for a few years, the interested parties are still keen to develop the alternatives, such as phase-change memory (PCM), ferroelectric and magnetic nonvolatile memories. However, apart from the development-stage competition, there is also the struggle to create a commercial product from a laboratory proof-of-concept.

Nanocrystals

Another direction is to further vary the existing Flash memory structure or materials. One such variation is to use silicon nanocrystals (Si-NCs) for charge storage. The coverage of this technology in the 2007 edition of the ITRS was somewhat transitional. For a number of years, ITRS has covered development of Si-NCs in its 'Emerging Research Devices' (ERD) section. In 2007, however, the group responsible for this section proposed that Si-NC Flash memory be kicked out of its remit and upstairs into 'Processes, Integration, Devices and Structures' (PIDS). This indicates that the group believes that now is the time for NCs to move from laboratory prototypes to preparing for manufacture.

Discrete charge storage, such as Si-NCs, allows some important potential advantages, such as the use of a thinner tunnel oxide thickness to maintain the necessary charge retention of 10–20 years. A thinner oxide would allow lower program/erase voltages, resulting in less damage, greater endurance for write/erase cycles, and improved reliability. However, there are concerns that non-uniformity of NC deposition could lead to variation in the threshold voltages and therefore less ability to distinguish multi-bit charge states.

In November 2005, Freescale reported a 24Mbit silicon nanocrystal memory in a NOR configuration. STMicroelectronics is also working in this area in collaboration with the Consiglio Nazionale delle Ricerche-Istituto Microelettronica Microsistemi (CNR-IMM) in Catania, Italy, producing a 16Mbit NOR device in 2007 [2] with NC sizes averaging 3nm and 6nm in two samples and NC densities in the range $3–6x10^{11}cm^2$. The larger NCs induced a cell reliability weakness.

One important development has been the independent discovery by Freescale and by CNR-IMM/ST with

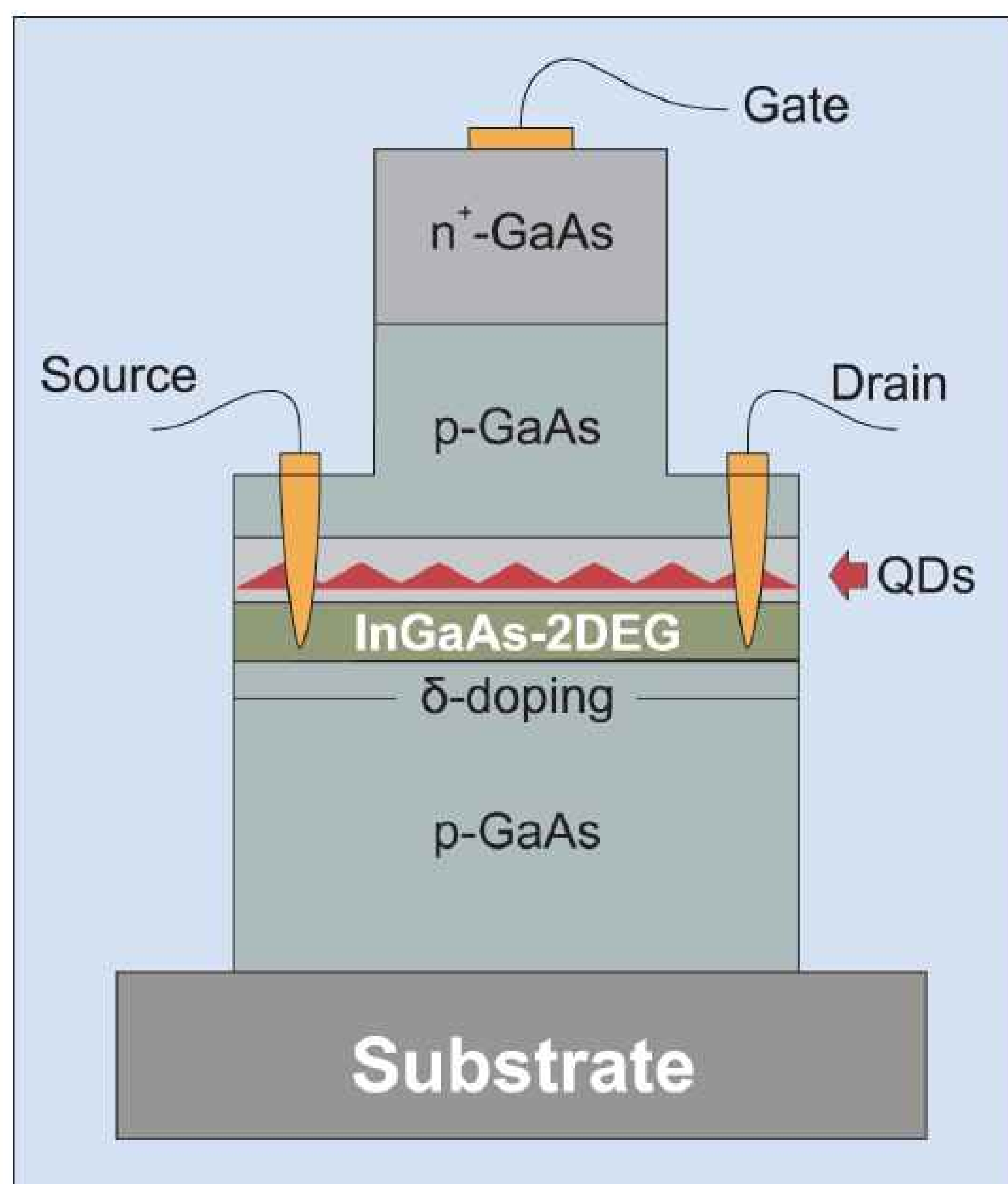

Figure 5. Schematic of quantum dot memory structure used by professor Bimberg's group at TU Berlin.

France's CEA-Leti [3] in the last couple of years that the deposition process is not completely random.

"During the formation of Si-NCs through chemical vapor deposition, a Si-free denuded zone forms around each dot, due to the diffusion of Si atoms on the substrate," says Rosaria Puglisi of CNR-IMM. "This denuded zone limits the number of available nucleation sites, the density of the dots, their size and relative distance, thus making the nucleation a non-random process."

Puglisi and her colleagues describe the state as being partially self-ordered. Compared with the assumption of a random Poisson distribution, the research finds that the presence of a denuded zone surrounding each dot leads to less dispersion from bit to bit of the surface coverage.

"This implies much more favorable projections concerning the ability of the nanocrystal memory concept to meet the scaling targets of future nodes," says Puglisi. "Experimental data on memory window dispersion fully support such a picture."

Another direction taken by CNR-IMM/ST with Belgium's IMEC, CEA-LETI and the University of Pisa has been to put SONOS and silicon nanocrystal charge storage layers into double-gated and triple-gate FinFET structures (Figure 4) [4]. In mainstream CMOS logic, multi-gated transistor devices are designed to reduce 'short-channel effects', allowing better electrostatic control of the channel. Double-gate devices are expected in about 2011. For Flash, such improved control could be used to improve threshold voltage windows. The CNR-IMM/ST research is aimed at pushing the scaling of Flash memories beyond the 28nm technology node (2015 and after). It is believed that channel lengths down to 10nm may be possible.

Enter III-Vs and quantum dots

While most Flash research focuses on increasing bit densities, some are looking to increase operating speeds towards those of magnetic hard drives and even DRAM. Although this work is still predominantly carried out in silicon, the research group of professor Dieter Bimberg at the Technische Universität Berlin (TU Berlin) is looking to use the much broader capabilities of III-V semiconductor materials for creating near-DRAM performance.

One of the researchers in this group, Martin Geller, comments: "The big advantage of III-V semiconductors, in contrast to group IV materials like Si or Ge, is on the one hand the much better crystalline quality of heterostructures, and on the other the possibility of 'band-structure engineering'. For band engineering in Si-based electronics, you are limited by the small number of material combinations of Si, SiO_2, SiN, Ge and a limited number of fixed band offsets. III-V materials have many more combinations — such as GaAs, InAs, GaSb, InSb, AlAs,... — not to mention more complex systems like InGaAs, AlGaAs,... These offer much greater potential for solving the long-standing problems of Flash, such as endurance and write/erase times. So, while at the moment every memory proposal that is not based on Si is viewed with skepticism, this may change in the future."

Dieter Bimberg at the Technische Universität Berlin is looking to use the much broader capabilities of III-V semiconductor materials for creating near-DRAM performance

Defects will become a big problem in future memory devices based on 'incoherent' Si or NC material. At highly shrunk feature sizes, leakage currents through defect states become a major obstacle. Here, self-organized coherent materials will be at an advantage. Professor Bimberg reports that his group already has some connections with industry partners.

TU Berlin has developed quantum dot (QD) memory structures with writing times of the order of 10ns

Along with researchers at Istanbul University in Turkey, TU Berlin has developed quantum dot (QD)

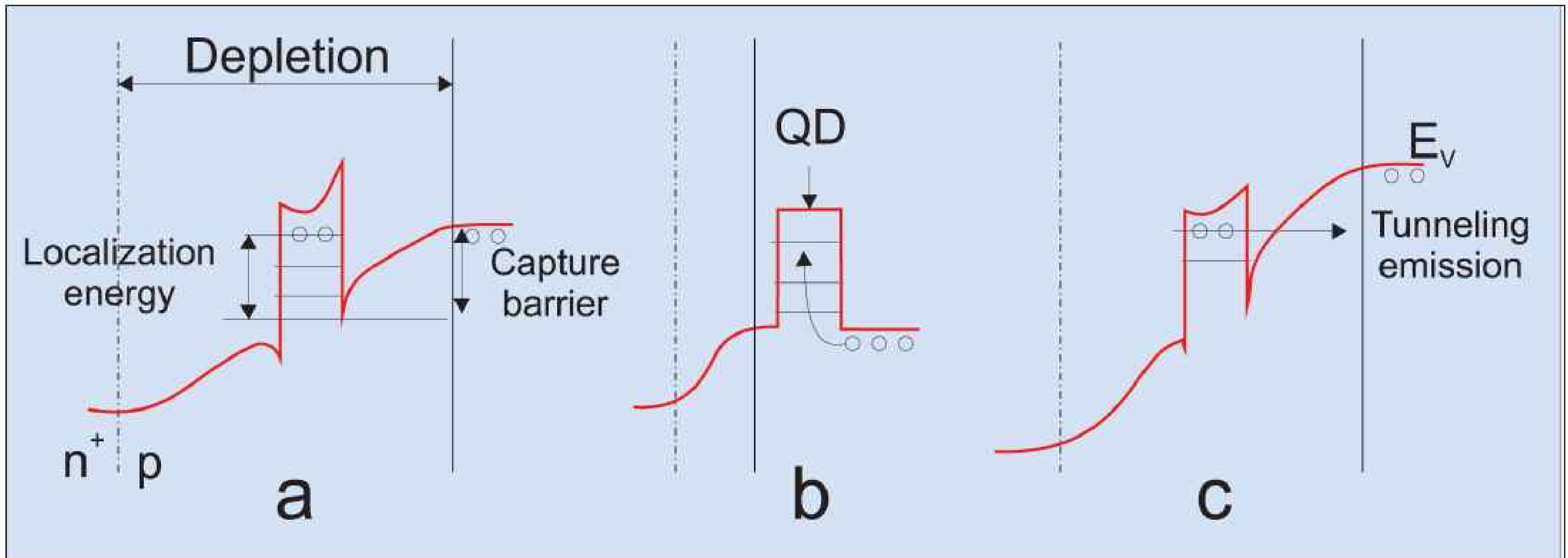

Figure 6. Structure of valence band for biases designed for storage (a), write of hole charge into quantum dot (b), and erase (c). Holes 'float' up to reduce electron energies.

memory structures (Figure 5) with writing times of the order of 10ns [5]. One QD structure had a write time as small as 6ns, while another arrangement was somewhat slower (14ns). Although these devices are limited by the experimental setup and by the cut-off frequencies of RC parasitics, the researchers hope to be able to use the concept to achieve even faster memory write/erase, based on picosecond charge-carrier relaxation times. These speeds compare extremely favorably with Flash's millisecond scale.

The fastest QD memory device was constructed from indium arsenide (InAs) embedded in a p-doped gallium arsenide (GaAs) layer, while the other was based on gallium antimonide (GaSb) embedded in GaAs. QDs are islands of the order of 10–50nm in diameter that 'grow' into more or less pyramidal shapes by self-organization due to the strain caused by the different lattice constants of the materials. An n-doped GaAs layer is then deposited on top of the QDs. The QDs measure about 15nm across. The estimated possible storage density of such devices is about 1TeraByte (1000GB)/inch^2 or ~160Gbits/cm^2.

One difference from normal Flash is that in the new QD device the barrier height that retains the charge is varied by using electrical means. A further difference is that, by growing the dots in the depletion region of a pn junction, it is holes that are stored in the QDs. By changing the bias on the depletion region, one can easily vary the barrier height, enabling either retention or the insertion of charge into the QD (Figure 6). Charge removal (erase) is achieved by using tunneling. The read mechanism is similar to that of Flash memory.

One difference from normal Flash is that in the new QD device the barrier height that retains the charge is varied by using electrical means

There is also a hope that QD-based memory will be more robust in terms of memory retention and long-term reliability. A study by the group on hole emission from InAs/GaAs QDs across a $Al_{0.9}Ga_{0.1}As$ barrier, combined with a theoretical extrapolation to other QD systems, suggests that GaSb in an AlAs matrix could yield a storage time of more than 1 million years [6]. A different GaSb/AlAs composition may be needed to give a write speed faster than DRAM.

A study by the group on hole emission from InAs/GaAs QDs across a $Al_{0.9}Ga_{0.1}As$ barrier, combined with a theoretical extrapolation to other QD systems, suggests that GaSb in an AlAs matrix could yield a storage time of more than 1 million years

It is also possible to produce Si/Ge QDs, but TU Berlin researchers found the system to have a storage time of the order of microseconds at room temperature, and no band-structure engineering is possible to improve this significantly. Hence, Bimberg's group has dropped investigation of group-IV QDs for this application.

References

1. ITRS 2007 at www.itrs.net
2. Gerardi et al, IEEE Transactions on Electron Devices, Vol. 54, p1376, 2007
3. R.F. Steimle et al., Microelectronics Reliability Vol.47, p585, 2007; R.A. Puglisi et al, J. Appl. Phys. Vol. 100. p086104, 2006
4. Lombardo et al, International Electron Device Meeting, 2007; this work was part of a European IST-NMP project, 'FinFET structures for FLASH devices (FinFLASH)' (2005-2007) www.imm.cnr.it/imm/progetti/projects/FinFLASH/index.html
5. Geller et al, Appl. Phys. Lett. Vol.92, p092108, 2008
6. Marent et al, Appl. Phys. Lett. Vol.91, p242109, 2007