Nanodevices: from novelty toys to functional devices – an integration perspective

Dinesh Pamunuwa[1] and Roshan Weerasekera[2]

[1]Lancaster University
Centre for Microsystems Engineering
Lancaster LA14YW, UK
dinesh@lancaster.ac.uk

[2]Royal Institute of Technology (KTH)
IMIT/LECS
Electrum 229, SE-164 40 Kista, Sweden
roshan@imit.kth.se

ABSTRACT

This paper looks at the prospects of continuing Moore's law into the deca nanometer regime using novel technology that has been recently proposed in the literature. It reviews some key advances in nanoelectronics, and provides an integration perspective for the ultimate goal of terascale integration. Issues from physical level circuits to system level architectures are discussed.

1. INTRODUCTION

Underlying the astonishing advances in information technology over the past 4 decades has been the ability of process engineers to continuously scale down the minimum feature size that can be printed on a semiconductor wafer through lithography. Hence the size of the Metal Oxide Semiconductor Field Effect Transistor (MOSFET), the workhorse of contemporary microelectronic circuits, has shrunk by several orders of magnitude during this period. The benefits of scaling can be analysed by examining the operation at the heart of all digital computation: the binary switching transfer implemented by the MOSFET.

MOSFET scaling gives us two obvious benefits; firstly, smaller unit sizes means that a greater number of units can be fitted in a given area. This increases the computational capacity (number of binary switching transfers that can be accomplished) for the same area from one technology generation to another, and is often presented as a saving in cost per transistor. Secondly, scaling reduces the physical length of the channel, which in turn leads to a faster switching speed, as to a first-order, the drift velocity of the electrons (and holes) within the channel of the MOSFET is proportional to the applied field.

These obvious benefits apart, scaling of the MOSFET sizes is accompanied by voltage scaling, as reductions in oxide thicknesses means that a constant electrical field (electrical stress) in V/cm can be maintained with a reduced voltage. This gives us a third, equally important benefit in that the energy of a binary switching transfer is proportional to the square of the voltage. Hence, for example, halving the supply voltage results in a 4-fold reduction in the energy.

The potential benefits of scaling were recognised in an oft-cited paper by Gordon Moore [1], who worked with William Shockley in the first ever high-tech start-up company of Silicon Valley, and went on to co-found Fairchild Semiconductor and Intel. He predicted that the number of transistors that could be integrated onto a single die (unpackaged chip) would double every year. He made this prediction from two data points, but it has proved to be remarkably accurate, as evidenced by the plot in Figure 1 of millions of transistors in lead microprocessors over the years. This seeming "something for nothing" violation of the free-lunch principle can be summed up in the incredulity with which Carver Mead, another pioneering figure in modern microelectronics regarded his own calculations in the fragment reproduced from [2] below:

> *"..I had been thinking about Gordon Moore's question, and decided to make it the subject of my talk. As I prepared for this event, I began to have serious doubts about my sanity. My calculations were telling me that, contrary to all the current lore in the field, we could scale down the technology such that everything got better: the circuits got more complex, they ran faster, and they took less power -- WOW! That's a violation of Murphy's law that won't quit! .."*

One specific question that Mead considered was the impact quantum tunnelling would have upon scaling, and his calculations revealed no significant effects for feature sizes that would enable integration of $10^7 \sim 10^8$ devices/cm^2. Many researchers considered the length of 1μm a significant barrier. That milestone has long gone, and Intel have a stable 65

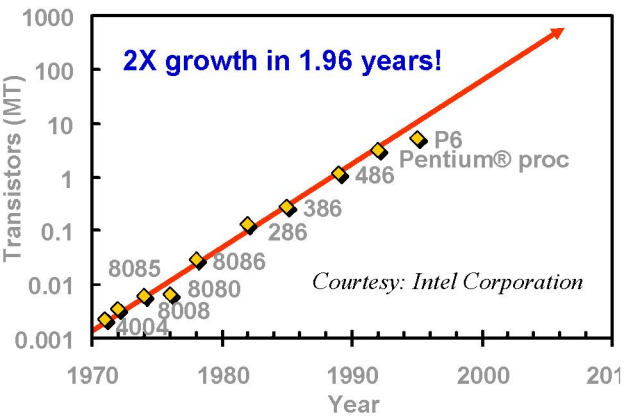


Figure 1. Growth of Transistors in Intel Microprocessors

1-4244-0322-7/06/$20.00 ©2006 IEEE

nm technology for microprocessors and a 45nm technology for SRAM. The end of the roadmap [3] for scaling of the conventional MOSFET is however in sight, due to a variety of reasons ranging from technological to fundamental.

When the dimensions of a MOSFET are scaled down, the area of the gate oxide is reduced. Since the gate oxide acts primarily as a capacitor, the thickness has to be reduced to compensate for the reduction in area. This means the voltage level must also be reduced to avoid material breakdown. Since the electron thermal voltage, kT/q (where k is Boltzmann's constant, T the temperature in kelvins and q the charge of an electron in coulombs) is a constant at room temperature, the ratio between the operating voltage and the thermal voltage inevitably shrinks. This leads to higher source-to-drain leakage currents stemming from the thermal diffusion of electrons. At the same time, when the gate oxide has been scaled to a thickness of only a few atomic layers, quantum-mechanical tunnelling gives rise to a sharp increase in gate leakage currents [4]. The effects of these fundamental factors on CMOS scaling manifest themselves in increased leakage current and reduced performance (drive current). Additionally, there are numerous technological hurdles to solve in scaling beyond the 24nm node [3].

To continue reaping the benefits of scaling, we need to gradually evolve to true quantum nano devices that work on quantum principles rather than continue to battle against these effects in the conventional MOSFET structure. The rest of this paper will outline some promising devices that hold out the potential of being a mainstream technology in the long term to rival Complementary Metal Oxide Semiconductor (CMOS) technologies, and identify key challenges from an integration perspective.

2. PHYSICS OF QUANTUM NANODEVICES

This section is a simplified explanation of the physics underlying the operation of some important quantum nano devices. Quantum nanodevices can be categorised into solid-state and molecular devices. Although lithographically fabricated solid-state nanodevices are unlikely to be the solution to extending the roadmap as discussed in section 3, it is insightful to understand the physical operation of key solid-state nanoelectronic devices; the operation of relevant molecular devices can be understood qualitatively using similar arguments.

As explained in [5], the essential structural feature that all solid-state nanoelectronic devices have in common is a small "island" composed of semiconductor or metal in which electrons may be confined. The extent of confinement of electrons in the island (i.e. whether having 0, 1, 2 or 3 classical degrees of freedom) defines the categories of quantum dots (QDs), resonant tunnel devices (RTDs) and single-electron transistors SETs. The composition, shape, and size of the island gives the different devices their distinct properties. Controlling these factors permits the designer of the device to employ quantum effects in different ways to control the passage of electrons on to and off of the island.

Two essential quantum mechanical effects are exhibited by electrons confined to nanometer-scale islands between closely spaced potential energy barriers. First, each electron's energy is restricted to one of a finite number of one-electron energy levels (quantum states with discrete, "quantised" energies). The smaller the distance between the barriers (i.e., the smaller the island), the more widely spaced in energy are the levels for the electrons in the potential well between the barriers. Second, if the potential barriers are thin enough (approximately 5–10 nm or less, depending on the height of the barriers), electrons occupying energy levels lower than the height of the barrier have a finite probability of "tunnelling" through the barrier to get on or off the island. Tunnelling is a consequence of the wave-nature of particles (in this case the electron). The wavefunction decays exponentially into the barrier, but if the barrier is thin enough, there is a non-zero probability of the electron ending up on the other side of the barrier.

There are two energies associated with an electron tunnelling onto the island as shown in Figure 2.a. The symbol $\Delta\varepsilon$ represents the spacing between two energy levels in a potential well and is the excitation energy, being the energy required to excite an electron to make the jump to the next level. There is also an energy U, the charging energy, associated with the electrostatic force which repels an electron trying to enter a space already occupied by N electrons. Thus the difference in energy between the lowest quantum state for N electrons and that for N+1 electrons for an island is $\Delta\varepsilon + U$. When a bias voltage is applied across the island, it induces mobile electrons in the conduction band of the source region to attempt to move through the potential well in the island region to get to the region of lower potential in the drain region. The only way for electrons to pass through the device is to tunnel on to and off the island through the two high potential barriers that define the island and separate it from the source and the drain. But tunnelling can occur and charge can flow toward the drain only if there is an unoccupied quantum energy level in the well at an energy that matches one of the occupied energy levels in the source band. RTDs (and other categories of nanoelectronic devices) work by means of tuning the energy of the quantum states in the potential well on the island relative to the energy of the bands in the source and drain.

An example of this is diagrammed in Figure 2.a. Increasing the applied voltage bias across the device of (a) progressively lowers the energy of all the states in the well relative to the energies of the electrons in the source. This is shown in (b) and (c). When the bias potential is sufficient to lower the energy of an unoccupied one-electron quantum state inside the well to be within the range of energies for the source conduction band, the quantum well is said to be "in resonance" or "on," and current can flow onto the island and out to the drain. This is shown schematically in (c). Otherwise, current through the device is blocked—the device is "out of resonance" or switched "off," as in (b). This use of a variable applied bias to switch a tunnelling current on and off characterizes the operation of a two terminal RTD. Similar adjustment of the energy levels in the

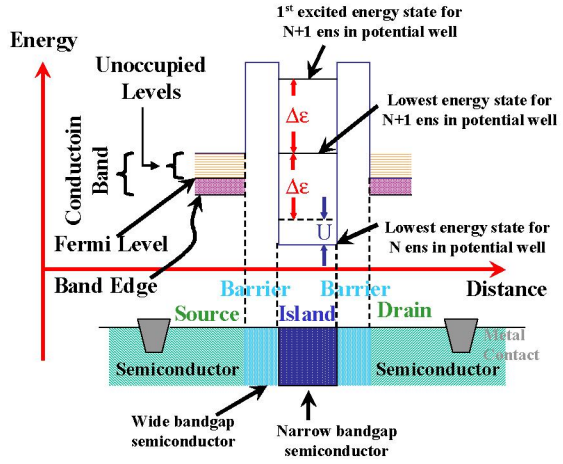


Fig. 2.a: Quantised energy states for a potential well

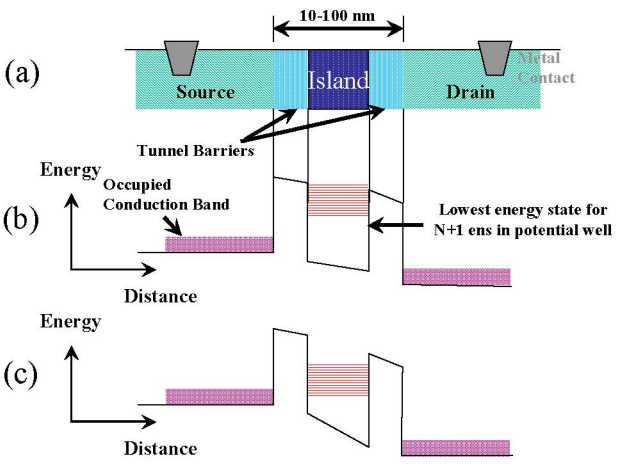


Fig. 2.a: Tuning Energy States for Controlled Tunnelling

Figure 2. Operation of Different Solid-State Quantum Nanodevices (source: [5])

potential well relative to those in the source also can be achieved by varying the voltage on a third (gate) terminal, rather than the voltage on the source. Such a three-terminal configuration results in an RTT, where a small gate voltage can control a large current across the device. Thus, an RTT can perform as both switch and amplifier, just like a conventional MOSFET.

QDs are constructed with islands that are short in all three dimensions, confining the electrons with zero classical degrees of freedom—i.e. electronic states are quantised in all three dimensions. The dot-like island may be made of either metal or semiconductor. Making an island short in all three dimensions leads to widely spaced quantum energy levels for an electron on the island. The charging energy is also large, because there is no way for a pair of electrons to get far from each other. This results in a different IV curve than an RTD.

An SET is always a three-terminal device, with gate, source, and drain, unlike QDs and RTDs, which may be two-terminal devices without gates. Also, an SET's island has no very short dimension and no very long one either. The island is usually made of metal and emphasizes U over ΔE, a defining characteristic of the energetic profile of the SET.

An SET switches the source-to-drain current on and off in response to small changes in the charge on the gate amounting to a single electron or less. In order to control the number of electrons on the island, a metal gate electrode is placed nearby. A sufficient increase in the voltage of the gate electrode induces an additional electron to tunnel onto the island from the source. The extra electron soon tunnels off onto the drain. This double-tunnelling process repeats millions of times a second, creating a measurable current through the island. As the gate voltage is increased, the number of electrons on the island stabilises at one more than previously, resulting in a drop in current. Further increases result in the same behaviour, so that multiple on-states can be identified at different gate voltages. For a more detailed explanation of the phenomena described briefly here, the interested reader can refer [5]-[8].

3. PERSPECTIVES FOR FUTURE INTEGRATION

The main premise of moving to a fundamentally different technology than CMOS is greater integration with acceptable reliability. One of the key challenges lies in the fabrication of the material islands, the requisite sizes of which can be quantified by a simple analysis. The change in energy for a tunnelling event is quantised as explained earlier, and can be described by

$$E_c = e^2/2C \qquad (1)$$

where C is the total island capacitance. This is identical to the charging energy associated with a capacitor. For this energy to be significant, it has to be much greater than the thermal energy (where k_B is Boltzmann's constant and T is

$$e^2/2C \gg k_B T \qquad (2)$$

the temperature in kelvin) as otherwise thermal fluctuations will mask it.

Equation (2) establishes a straightforward relation between the capacitance and hence the geometry of the island, and the operating temperature. As an example, for an island size on the order of 1μm, a junction area of approximately 100nm x 100nm with an oxide layer thickness of 1nm, the capacitance is on the order of 1fF. This corresponds to a temperature of about 1K. However for reliable operation in a digital environment, the charging energy should be larger than the thermal energy by a factor of around 100. At room temperature, the corresponding island size is approximately 1nm. One of the major challenges for the use of single electronic devices in integrated

circuits (ICs) is the *reproducible* fabrication of circuits with such small features.

3.1 Fabrication of Single Electronic Devices

The current state of the art in lithography allows feature sizes of around 20nm to be printed on a Si wafer, and there are techniques such as electron-beam lithography which promise smaller dimensions, but these are presently uneconomical [3]. Most innovative bottom-up techniques which have resulted in measurable single-electron behaviour suffer from the disadvantage of unpredictable device parameters, which precludes their use in ICs. Another possible technique is the use of scanning probes to manipulate material on a nano scale[1]. A major limitation of this method in IC fabrication is speed, since parallel fabrication of millions of devices is infeasible. To overcome the reproducibility problem and get the true benefit of single electronic phenomena, it appears that the most promising approach is to use nature's building block of 1nm dimension, the molecule.

Experimental observation of electron transport through molecules has been reported in the literature, the mechanism of which can be explained qualitatively as follows. A molecule has discrete electron orbitals, and if the molecule is long enough, the attachment of electrodes will not significantly affect these orbitals. Analogous to solid-state devices, there will be an energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). If the Fermi level of the electrodes falls between these levels, no current will flow in the absence of an applied voltage. An applied voltage will cause the HOMO or LUMO levels to be aligned with the Fermi level of each electrode in turn, and electrons will flow in two hops, passing from one electrode to the molecule, and from the molecule to the other electrode [9]. The physical analysis is very similar to that already described in Section 2, although some of the exact mechanisms of interaction are not completely resolved.

Although many molecular electronic devices have been demonstrated by physicists and chemists in laboratories, they mostly suffer from the same disadvantage of not being reproducible with any uniformity, accuracy or predictability in great numbers. One solution has been suggested by some promising recent developments in nanowire fabrication. The basic idea is to produce arrays of nanowires, and connect them with programmable nanodevices. Two key issues are, how to connect the nanowires to the nanodevices, and how to address the array.

The addressing problem can be solved by changing the property of the nanowire, such as its doping, *along its axis*, thus allowing some peripheral control circuitry comprising devices with a larger footprint to access it. The nanowires themselves can either be synthesized as free-standing wires such as in [10] or [11], and aligned into an array using flow techniques, or created as an ordered array using techniques such as nanoimprinting [12]. Two orthogonal nanowire arrays can be combined so that molecular nanodevices form the vertical connection [13], [14]. Since the devices can be programmed to be 'on' (short circuit) or 'off' (open circuit) this provides a programmable fabric, in this case an OR plane.

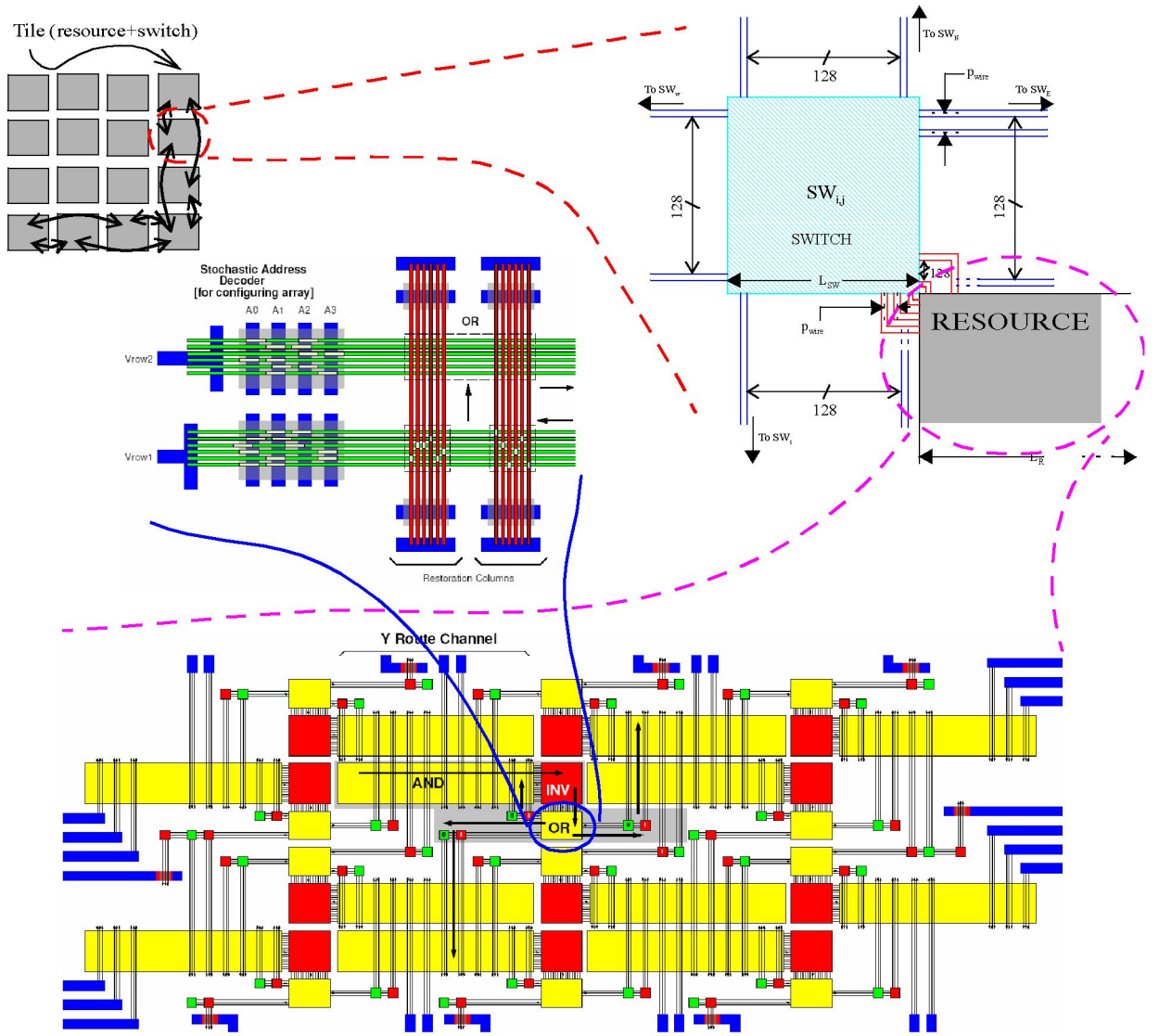
The combination of the nanodevices with the nanowires can be achieved by depositing a self-assembled monolayer (SAM) on one array, and then depositing the second array on it. A SAM is essentially a technique where a solution containing the required molecules is deposited and the excess is washed off. The molecules bond to the surface and hence remain.

3.2 Defect Tolerant Architectures

A second non-trivial problem in the use of single electronic devices is the problem of background charge. If any charged impurity is located at some distance from the island and this distance is comparable to the size of the island, a polarisation will result. This will affect all characteristics of the device, and essentially render that device unusable. Since there will always be an inescapable minimum impurity level in any fabrication process, this essentially means that some portion of the devices on a chip will be unusable, or in other words the *yield* will be lower than in a stable CMOS technology. The solution to this problem is more likely to be found in the architecture of the system than in technological improvements. There could also be transient variations due to leakage or external pertubations such as noise. Hence along with some percentage of the devices never working at all, a likelihood exists that some of the devices will work some of the time. This is potentially a bigger problem leading to intermittent system failures. Most fault-tolerant techniques for nanocomputing feature some sort of redundancy [15]. The basic idea is that a function is implemented many times, and the output is resolved through a majority gate. So for example, if the redundancy is a factor of three, and each block should output logic '1', but due to an error only two function correctly, the output is proportional to 2/3 rather than 1. By relaxing the threshold appropriately, the correct output can still be obtained. Many variations on this theme exist, including a multiplexing scheme which allows functional circuits to be built out of building blocks with a failure rate of around 1%, but requires redundancy factors on the order of 10^3~10^4. Generally there is a clear trade-off between area and reliability in all these approaches. The question of whether the levels of redundancy required in a specific molecular device technologies would allow favourable comparisons with CMOS in terms of integration density, power consumption and system bandwidth is unresolved.

Also, there needs to be a clearly defined interface to CMOS from the nanoscale building blocks in a physical sense and in a methodological sense. Most bottom-up fabri-

1. One of the best known applications of scanning probe techniques is in memory, such as the Millipede initiative by IBM, which uses heated Atomic Force Microscope (AFM) tips to make tiny indentations in a polymer. the absence or presence of an indentation represents 1s and 0s.



Top left is a folded torus network-on-chip topology from [18] which represents a way of managing on-chip interconnect woes; each tile represents a resource and switch, where the resource implements custom logic, while the switch communicates with the on-chip network, and provides the network interface; a tile is expanded into a switch and resource on the top right; the resource is implemented as a nano PLA from [14], but it can be any nanoscale implementation in general; the OR plane is expanded in the middle left picture, again as specified in [14].

Figure 3. A possible architecture for terascale integration

cation techniques are feasible only for 2-terminal devices, most of which are non-restoring. CMOS provides a simple regenerative technique for interblock communication. More generally, CMOS will provide the means of interfacing to other technologies and the rest of the world, and the architecture should reflect this.

From a methodology point of view, the past 30 years has seen the development of a well defined set of abstract models and integrated hierarchical design tool framework for CMOS, which has been essential to complex multi million device IC design. Novel technologies which provide very small building blocks need to be encompassed to a traditional CMOS design flow, to take advantage of this well-established framework.

Another crucial point in the architectural design stems from the properties of wires. On-chip wires are commonly made of Al or Cu, and contrary to devices, become slower the more they are scaled. The reason is, they become highly resistive and stop behaving like equipotential regions. Rather, signal propagation is governed by the diffusion equation, and associated with a wire is a time constant

$$\tau_w = R_w C_w \qquad (3)$$

Both R_w and C_w are essentially proportional to wire length, and hence τ_w is a quadratic function of wire length. This is one of the fundamental bottlenecks in on-chip communication [17], and emphasises the need for locality in system architecture. To mitigate this problem, scaling of wires is

carried out in a hierarchical manner, so that local wires are scaled aggressively, whereas global, chip-level interconnects are scaled much less.

Considering all these diverse requirements, one possible solution is outlined in Figure 3. As argued in [18], dealing with on-chip interconnection woes is made easier if there is regularity at the global level. Global communication within the chip is achieved by means of an on-chip network, whose parameters can be closely controlled. Each tile comprises a resource where custom logic or memory is implemented, and a switch, which provides the interface from the resource to the network. The resource can be implemented in any feasible technology, and the figure shows the architecture proposed in [14], called nano PLA. It must be emphasised that such a system level view is hypothetical, and considerable technical challenges remain. However it marries the best of both worlds, CMOS and nanotechnology, to utilise known strengths and mitigate the effects of known weaknesses.

4. CONCLUDING REMARKS

Some very exciting advances in state of the art technology has given the electronics design community the promise of integrating a trillion nanoscale devices on a single chip. The reliable building block that is the MOSFET suffers from short-channel degradations which are manifestations of quantum mechanical phenomena at channel lengths around 25nm. Therefore it makes sense to explore alternate devices which operate on these very principles, rather than in spite of them. From an analysis of the fundamental physics, and taking into account realistic technological capabilities, one of the best perspectives for reliable assembly of nano devices on a large scale uses nanowires, with the devices forming a vertical connection between two arrays. Several research groups have demonstrated potential approaches based on this concept.

An important caveat in building a terascale system is the presence of a CMOS interface, at difference hierarchical levels within the system itself, and also as a connection to the outside world. This paper has briefly investigated a possible architecture that satisfies this requirement, using known methods and also hypothesizing solutions based on advances reported in the literature.

5. ;REFERENCES

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, pp. 114-117, 1965.

[2] T. Hey and P. Walters, *The new Quantum Universe*, CUP, 2nd ed., p. 187, 2003.

[3] International technology roadmap for semiconductors (ITRS). [Online]. Available: http://www.itrs.net/Links/2005ITRS/Home2005.htm.

[4] Y. Taur, "CMOS design near the limit of scaling", *IBM J. Res. & Dev.* vol. 46 no. 2/3 March/May 2002.

[5] Goldhaber-Gordon et al., "Overview of Nanoelectronic Devices," *Proc IEEE*, vol. 85, pp. 521-540, 1997.

[6] K. K. Likharev, "Electronics below 10nm," *Nano and Giga Challenges in Microelectronics* (Amsterdam: Elsevier), pp. 27-68, 2003.

[7] K. K. Likharev, "Single-electron devices and their applications," *Proc. IEEE*, vol. 87, pp. 606-632, Apr. 1999.

[8] G-L. Ingold, "Charge tunneling rates in ultrasmall junctions," *Single Charge Tunneling*, NATO ASI Series B, vol. 294, (Plenum Press, NY), pp. 21-107, 1992.

[9] A. Nitzan and M. A. Ratner, "Electron Transport in Molecular Wire Junctions," *Science*, vol. 300, pp. 1384-1389, May 2003.

[10] Y. Cui et. al., "Diameter-controlled synthesis of single-crystal silicon nanowires," *Applied Physics Letters*, vol. 78, no. 15, pp. 2214-16, 2001.

[11] P. L. McEuen, M. Fuhrer, and H. Park, "Single-walled carbon nanotube electronics," *IEEE Transactions on Nanotechnology*, vol. 1, no. 1, pp. 75-85, Mar. 2002.

[12] M. D. Austin et. al., "Fabrication of 5 nm linewidth and 14 nm pitch features by nanoimprint lithography," *Applied Physics Letters*, vol. 84, no. 26, pp. 5299-5301, Jun. 2004.

[13] D. B. Strukov and K. K. Likharev, "CMOL FPGA: a reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices," *Nanotechnology*, vol. 16, pp. 888-900, 2005.

[14] A. DeHon and M. J. Wilson, "Nanowire-based sublithographic programmable logic arrays," in *Proc. ISFPGA*, Feb. 2004, pp. 123-132.

[15] K. Nikolic, A. Sadek, and M. Forshaw, "Fault-tolerant techniques for nanocomputers," *Nanotechnology*, vol. 13, pp. 357-362, 2002.

[16] J. von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," *Automata Studies*, eds. C. E. Shannon and J. McCarthy, Princeton NJ, pp. 43-98, 1955.

[17] D. Pamunuwa, L-R. Zheng, and H. Tenhunen, "Maximising throughput over parallel wire structures in the deep submicrometer regime," *IEEE Transactions on VLSI*, vol. 11, no. 2, pp. 224-243, Apr. 2003.

[18] D. Pamunuwa et. al., "A study on the implementation of 2-D mesh-based networks-on-chip in the manometer regime," *Integration*, vol. 38, pp. 3-17, 2004.