**ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY**
Volume **12**, Number 3, 2009, 317–338

# Nanoscale Electron Devices: Difficulties, Granularities and Applications

Daniel FOTY

*Gilgamesh* Associates LLC, 452 Black Mountain Rd., Fletcher VT, USA
E-mail: dfoty@sover.net

**Abstract.** Some critical (but little-recognized) aspects of nanoscale electron device technology are considered. According to history, nanoscale electron devices will have to address some convergence of an abundance and a scarcity. A major impediment to progress is the emergence of a variety of granularities – a problem as much intellectual as material.

## 1. Introduction

Following several years of hype and hope, general disillusionment with "nanotechnology" appears to be setting in; this is best exemplified by the "Eighteen-billion-dollar pair-of pants" [1] jibe, which has become symbolic of the original hubris surrounding nanotechnology.

In October 2007, a special invitation-only review colloquium, under the auspices of the IEEE Electron Devices Society and IEEE Solid State Circuits Society, was held in Sinaia, Romania [2] (this author was among the invited participants [3]). The overall mood was extremely gloomy, as the original hopes seem to have catastrophically smashed themselves to bits on the rocks of scientific and engineering realities; one participant even went so far as to quote the enduring truth of his own published 2003 observation that "There is more hype than reality" [4]. However, it is when things are at their gloomiest that we can examine things soberly and make sensible evaluations.

This paper will evaluate prospects for electron devices at "nanoscale" dimensions. Contrary to much of the hype surrounding this topic, it is found that our best "nanoscale electron devices" may already be with us. This situation is contrasted against a variety of serious technological problems that confront "very small" electron device technology.

## 2. Philosophy

To begin, we must consider two philosophical points; this basic understanding is necessary for the evaluation of nanotechnology – and for the evaluation of its prospects for genuinely revolutionary changes in technology.

Way back in 1996, the noted technology analyst George Gilder made an important observation: "Every economic era is based on a key abundance and a key scarcity" [5]. This is a very useful guidepost; as we attempt to evaluate nanotechnology, we must look for confluences of abundance and scarcity where nanotechnology can be decisive.

We can also note the 1971 observation of the late Czech-American engineer and philosopher Petr Beckmann: "In a healthy society, engineering design gets smarter and smarter; in an [unhealthy society], it gets bigger and bigger" [6]. Discontinuous leaps in engineering technology are based on deploying new knowledge – rather than on just "scaling things up." Based on Beckmann's comments, though, it has become easy to counter-assume that anything "smaller and smaller" must be "smarter and smarter;" this has been particularly true with the stunning shrinkage of microelectronic dimensions over the past several decades. So with regard to nanotechnology, we must carefully ask ourselves, "Is 'smaller and smaller" really 'smarter and smarter'? Or is 'smaller and smaller" really just the new 'bigger and bigger'?"

Thus, we can see that the main challenges of nanotechnology are more intellectual than material. New thinking is required; to make full use of nanotechnology, "we must rise with the occasion" [7].

## 3. Granularities

A critical (but little-recognized) problem in nanotechnology is the emergence of "granularities" – things that have historically been treated in bulk and approximate ways but which can be so treated no longer.

Most particularly, the triumph of "digital" technology has actually always been based on its fundamental coarseness. At the critical juncture of things-analog and things-digital, "analog" methods are efficient but ambiguous, while "digital" methods are unambiguous but inefficient. *Digital technology came to dominate because cheap bits and cheap MIPs overcame the inherent inefficiency.* This is one of our most salient present "fundamental abundances."

However, the inherent coarseness of digital methods permitted important physical details to be treated in a similarly coarse (but cavalier) manner. At nanotechnology dimensions, these more fundamental details begin to re-assert themselves.

A good example of this re-assertion is the seemingly simple and mundane concept of particle velocity under an electric field in a solid. The motion of charged particles in a solid (particularly a semiconductor) has been treated in an extremely simple fashion for decades; this level of simplicity has been viable due to the coarse and approximate nature of electron devices. However, at a "nanotechnology" level of granularity, the realities are much more complicated – as depicted in Figure 1. Unlike the case for

our "traditional" simplifications – where the velocity distribution could essentially be treated as monochromatic – we can see that the velocity distribution is much richer and much more complicated. We must discard our comfortable but outdated ways of approaching these problems if we are to properly deal with nanotechnology.

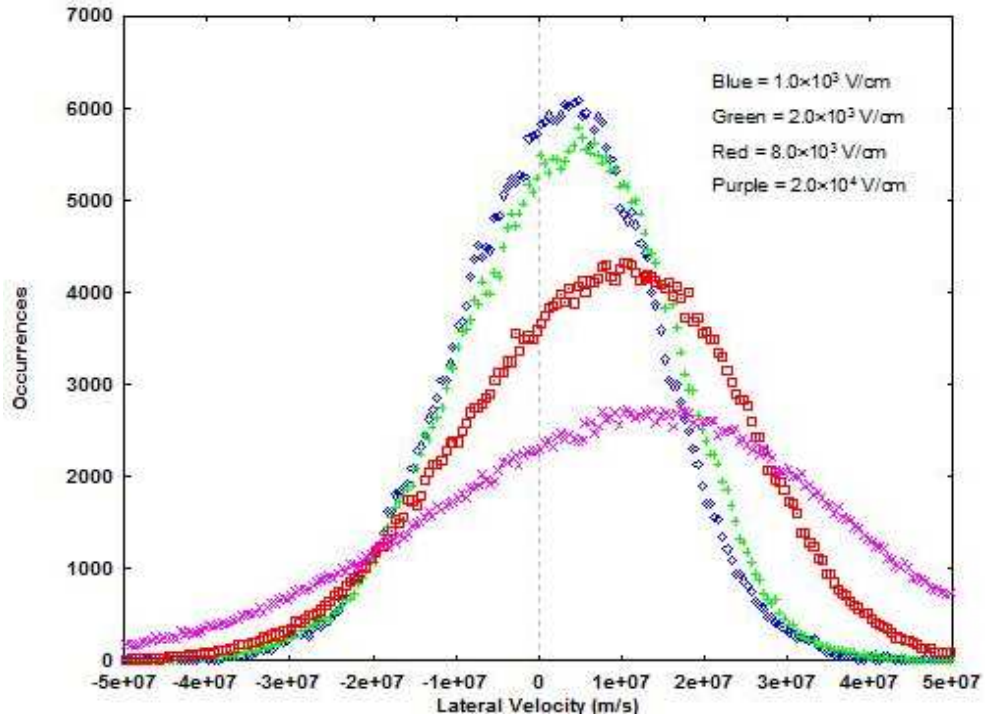


**Fig. 1.** The distribution of electron velocities in silicon under several different electric fields, calculated via Monte Carlo methods; the distributions are extremely broad, even for low electric fields.

More generally, if we look at all of our decades of accumulated intellectual machinery for understanding (and describing) the MOS transistor, we find that it is based on coarse granularities and "bulk" properties. That is, we have long assumed that items such as doping, oxide thickness, etc., can be completely described in "bulk" terms – both qualitatively and quantitatively. However, at nanoscale dimensions, these notions are not valid, as we can no longer think in terms of bulk material but must think on the atomic scale – as we have reached the limits of indivisibility. Not surprisingly, the challenges presented at these dimensions thus go far beyond those of the mere materials – these challenges are also intellectual, since there must be recognition that the old mental-models are inadequate to the situation.

A striking example of this dichotomy is provided by comparing the current-carrying properties of bulk copper (which is now used extensively in integrated circuit interconnect metallurgy) and those of carbon nanotubes. It has been known for some

time [8] that carbon nanotubes have much higher current-carrying capacity; copper falls apart (literally) at a current density of about $10^6$ A/cm$^2$, while carbon nanotubes can survive to the much higher current density of $10^9$ A/cm$^2$. However, as a material, copper can indeed be treated in a "bulk" fashion in this use – copper can be deployed "in bulk" into real applications, and the areal property holds at any presently-relevant dimension. In contrast, carbon nanotubes *cannot* be "scaled up" into a bulk material – the gaudy current-carrying numbers are a mathematical artifact of tiny individual nanotubes, and that property cannot be scaled up to the dimensional sizes needed for real-world applications.

A second critical granularity is provided by the temperature – as embodied by the thermal voltage kT/q. Since Boltzmann's constant $k$ and the electron charge $q$ are fundamental physical quantities, they obviously cannot be changed. For "real world" scenarios, the temperature is basically fixed; thus, we are left with a thermal voltage of 26 meV.

At the nanoscale, the small dimensions also demand very small electric fields – fields at essentially the same levels that have historically been used in transistors. This naturally results in a demand for very low voltages – and those voltages quickly begin to draw perilously close to the thermal voltage. This is obvious, but once again the challenges presented are subtle; all of our methods of understanding and using electron devices have been built around an assumption of working voltages that are considerably larger than the thermal voltage. New methods of approaching this situation are required.

Interestingly, this anchored limit imposed by the thermal voltage has already had a drastic impact on the nature of MOS technology – and quietly has for more than a decade and a half. This particular issue will be considered in detail later.

## 4. Salient Points and Application Targets

We can now cast a wider net to see how nanotechnology might fit into the more general cast of contemporary technological challenges.

Earlier, an observation was noted that: "Every economic era is based on a key abundance and a key scarcity." Gilder [5] made that observation in 1996 – and noted that in electronic engineering, power and transistors are abundant and are used profligately; the latter observation comports with the aforementioned notion that digital technology has succeeded because very expensive transistors – and thus bits and MIPs – allow the inherent inefficiency of digital methods to be overcome. The key scarcity, however, is bandwidth. Thus, the key abundances (power and transistors) are used to try to address the key scarcity (bandwidth). Twelve years on, we find that the situation trenchantly identified by Gilder in 1996 still holds today; in fact, the situation has worsened, as demand for bandwidth has increased, both space and time have become more scarce, and methods of trying to "create" bandwidth achieve little while consuming more and more power.

Writing in 1996, Gilder looked ahead to a new paradigm – one in which bandwidth would be abundant, but in which *power* would be used sparingly (thus serving as the

"key scarcity"). This shift hasn't happened yet, despite being even more needed than it was back in 1996; the dizzying growth in (increasingly-capable) portable devices is demanding this change – yet technology has not delivered.

Thus, a driving force for "nanotechnology" should be the need for much more bandwidth at much lower power consumption. The required improvements are not minor, but in the orders-of-magnitude category – data rates of terabits-per-second and petabits-per-second will be required. On that last count, we can note with care that compression technology is one of the most notable attempts to "fake" bandwidth by using power-hungry digital processing; however, the quality of high-definition (HD) video displays has become so good that the inferior quality of compressed video is discernable to (and unacceptable to) the user. The impending proliferation of HD video demands *uncompressed* distribution – and that will require network data rates of petabits per second.

We can also note that wireless bandwidth in particular presents a special challenge. The need for gigabit per second wireless communications capability has long been known, but has never been delivered – largely because various attempts have always (blindly) relied on baseband processing methods that consume too much power. High power consumption is of course particularly unacceptable for portable devices.

This leads us straight into what has been the biggest impact to date of nanotechnology – the incredible growth in storage capacities, and the extreme compactness of that storage. In October 2007, the Nobel Prize for Physics was awarded for the fundamental work on "giant magnetic resonance," which opened the door to the recent runaway expansion in disk storage and concomitant shrinkage in disk size. With more storage and an increase in the portability of that storage, we are now seeing a "data diaspora," in which there is not only more data – but in which that data is more spread out and found in more places.

A final salient factor is technology cost. MOS technology has an interesting dual-life on this count. On the one hand, CMOS technology is inexpensive and ubiquitous; *it is the low cost of CMOS that has been the prime driver for the long-term dominance of that technology in the world today.* Today, we have reached the point where the cost of a MOS transistor is about 100 *nanocents* ("1 cent" = \$0.01) [9] – or about *one* ***nano****dollar (!).*

On the other hand, the structural costs involved with MOS technology provide a frightening picture. The cost of a fabrication facility has been growing exponentially *for more than twenty years.* At this time, the ground-up cost of a new fabrication facility is approaching some \$15 billion; if present trends were to hold, in another twenty or thirty years the projected fab cost will approach the level of total world GDP – something that obviously cannot happen. In addition, it must be noted that in CMOS technology, the cost of mask sets has become equally frightening; costs have reached \$800,000 at the 90nm technology node, \$1.2 million at the 65nm node, and greater than \$2 million at the 45nm node. Naturally, these exorbitant costs serve to greatly increase risk – and to thus discourage investment and innovation.

The cost/benefit trade-offs and the economics of the contemporary semiconductor industry present a very tight squeeze – one that only continues to tighten. *Nanotechnology will enjoy significant success when it offers an escape from this squeeze.* It

must offer something different – not just a further extension of what is already being done. It is always worth remembering that the "MOS miracle" has been based on something rather remarkable – incredible technological and economic value based on the combination of dirt and air. Nanotechnology must find some way to offer a similar sort of "miracle."

Thus, if nanoscale electron devices of a new form are to emerge and dethrone "King CMOS," several things must happen – as this excellent list [8] shows:

- It must be easier and less expensive to manufacture;
- It must be capable of very high current drive – high enough to drive interconnect reactances of essentially unlimited physical length;
- It must offer high integration ability – at or beyond what CMOS now offers (which exceeds 1 billion transistors per integrated circuit);
- It must offer excellent reproducibility;
- It must be reliable – reliable enough to meet the generalized benchmark of operation for at least ten years "in the field";
- It must offer very low cost, at least in the earlier-discussed range of effective device cost of the "nanodollar" sort;
- It must be very good at heat dissipation and transfer.

A superb explanation to summarize this is [8]:

> "*Everything about the new technology must be compelling, [while] further CMOS scaling must become difficult and not cost-effective. Until these two happen together, the enormous infrastructure built around silicon will keep the silicon engine humming.*"

Thus, the key "nanotechnology" challenge is to move beyond "more of the same," and to think more broadly; in particular, we need to move beyond the mindset that "progress" is solely a matter of making transistors smaller and smaller. In particular, what have traditionally been regarded as "network-level" principles can be pulled *inside* electron devices.

## 5. Semiconductor Electron Devices

As we evaluate possible opportunities for nanoscale electron devices, we must evaluate the situation with the most successful electron devices of all time – semiconductor electron devices. The goal is to discern the key factors in that success, and how those factors must also apply to possible nanoscale devices.

The key factors in the success of semiconductor electron devices have been **repeatability**, robustness, size, cost, speed, and the size of the resultant systems. **Repeatability** is something that has come to be taken for granted and that does not receive the recognition that it is due – when it is perhaps *the* most critical factor. Assured repeatability (from device to device) has been *the* defining deciding factor in favor of semiconductor electron devices.

The key implication, of course, is that any nanotechnology that does not measure up to the semiconductor standard of repeatability will fail; the vulnerability is particular glaring when is collides with the emergence of various "granularities" discussed earlier. As the granularities have begun to emerge even into mainstream silicon technology, this has begun to create a problem of controlling things that happen "rarely" [9]; this situation naturally applies to other forms of nanotechnology.

However, as noted above, "progress" in electron devices has come to be regarded solely as a matter of making transistors smaller and smaller. It is reasonable to ask if we can find new ways to get more mileage out of the silicon we already have – as these provide the real keys to nanoscale electron devices. This requires a cursory review of the key factors in the success of silicon electron devices

Semiconductor electron devices succeeded because they offered "clean" behavior – that is, they offered "chemical" interfaces rather than the "mechanical" (material) interfaces of their predecessors. In early semiconductor devices, germanium was the material of choice – largely due to the higher intrinsic carrier mobility that was available. However, this all changed with the development of the oxidation of silicon – as this offered the ability to easily create a controllable, **repeatable**, high-quality *in-situ* insulator. Thus was born MOS technology – which, as noted earlier, consists of a combination of dirt and air.

And thus a juggernaut was created. CMOS technology proved to offer great flexibility at reasonable cost, as well as superior logic stability – all in all, the property of being a very forgiving technology made CMOS a natural choice for practical applicability.

However, the most critical aspect of CMOS technology was indeed not material – but intellectual. This was the realization in 1972 by Carver Mead [10] that MOS was a unique technology for one particular reason – that it was amenable to *scaling*. What Mead realized – though this author will phrase the insight in a different way – is that for the MOS transistor, *as the dimension goes to the limit of zero, the voltage should to go the limit of zero with it*. This basically meant that the size reduction should be done in a way that kept the fields inside the device *constant* as the dimensions were sequentially reduced over time.

What is truly amazing is that this is something unique in technology – the situation was simplified so that dimensional shrinkages were rendered akin to photocopy reduction. Because a MOS transistor could be well-described by a few quantitative "bulk" characteristics, the scaling rules allowed the target values for the next generation to be computed and known *in advance of any process development*. This provided a winning one-two punch. First, with the process goals already known, these could be targeted directly – rather than by some murky set of results that might be amorphous or lack general applicability. Second, with the end goals known, work on the design of circuits and systems could begin *before the underlying process development was even finished – let alone ready for manufacturing*.

Scaling also combines nicely with another unique feature of the MOS transistor – that is, its dimensional flexibility. The basic nature of the MOS transistor allows its dimensions (length and width) to be adjusted at will *at the design level*. We have become so used to this ability that we have forgotten how unusual (and rare) it is.

Thus, it became relatively easy to produce MOS transistors in enormous quantities, to do so at low cost, and to continually improve their performance by rapid scaling to smaller and smaller dimensions. This is how transistors, bits, and MIPs became so cheap and plentiful that we could "waste" them in favor of other factors. Originally, the "waste" was mainly used to gain flexibility in designs. However, transistors have now become so inexpensive that they can be wasted in favor of programmability. Field-Programmable Gate Arrays (FPGAs) were originally developed for purposes of testing and prototyping – such as for checking the correctness of a large-scale logic design; however, they were too slow for real applications. Over time, with the rapid shrinkage of MOS technology, that last constraint has disappeared. FPGAs have now largely displaced Application-Specific Integrated Circuits (ASICs), as they are fast enough for use in "real" products. *FPGAs are technologically and financially superior, even though a large fraction (often as high as 75%) of the on-board transistors and logic gates are not even used.* The flexibility and rapid-development time are more than worth the waste. (Interestingly, the FPGA arena now also represents one of the few financially and economically bright parts of the entire semiconductor industry.)

As an aside, this discussion tells us something important about many recent efforts to development novel MOSFET structures, such as double-gate devices and FinFETs. Given the advantages described above for the basic MOS transistor, these novel structures would seem to fall in exactly the wrong spot – they are not different enough to offer a real upward operational change, yet they are different enough to be difficult to make, and therefore expensive.

Mead's original 1972 insight can be described as *constant-field scaling* – that is, as dimension and voltage are reduced at the same rate, the electric fields inside the MOS transistor remain *constant* even as the sizes grow smaller; it is this maintenance of the field strength that underpins the marvelous simplicity of "scaling."

However, CMOS technology did not follow that path. What actually happened (and what followed – which will be dealt with below) is depicted in Figure 2.

During the 1970s and into the 1980s, CMOS technology instead followed a path known as *constant-voltage scaling.* The reason for this choice was simple – there was intense resistance on the part of systems manufacturers to the notion of integrated circuit pin voltages changing every few years. Thus, the 5V power supply voltage – which CMOS had adopted for initial compatibility with the bipolar TTL technology that it replaced – was maintained for many years.

As can be seen in Figure 2, this evolution meant that the electric fields inside the MOS transistor were increasing over time as the dimensions shrank; *this is something that MOS technology does not naturally "want" to do.* Rather than preserving the basic inherent simplicity of the MOS transistor over time, the need to cope with the higher and higher field strengths required extensive and continuing modifications to the simple MOS structure – which greatly expanded the cost and complexity of the technology.

By the early 1990s, this situation had precipitated a wide-ranging crisis. The high voltages were inducing crippling problems with power consumption, and reliability problems due to the high fields were slowing the introduction of new technologies. In

addition, as noted above, the growing complexity of processes was causing them to become prohibitively expensive; this was also greatly slowing down the introduction of smaller-sized process generations – upon which the success of CMOS was built.

As Figure 2 also shows, this crisis finally caused the "re-discovery" of the original notion of constant field scaling. This was the basic theme that drove CMOS technology during the 1990s – the reduction of the voltage in proportion to the reduction of the dimension. Processes became simpler, the introduction of new process generations picked up speed, amortized costs decreased, and silicon technology diffused into more and more of the world.

As the lower-left corner of Figure 2 indicates, this return to the original Mead concept of constant-field scaling was projected to continue well into the $21^{st}$ century – even as dimensions dropped below 100 nm. CMOS technology has indeed rapidly breached the 100 nm barrier, and has by evolution become the first practical nanotechnology.

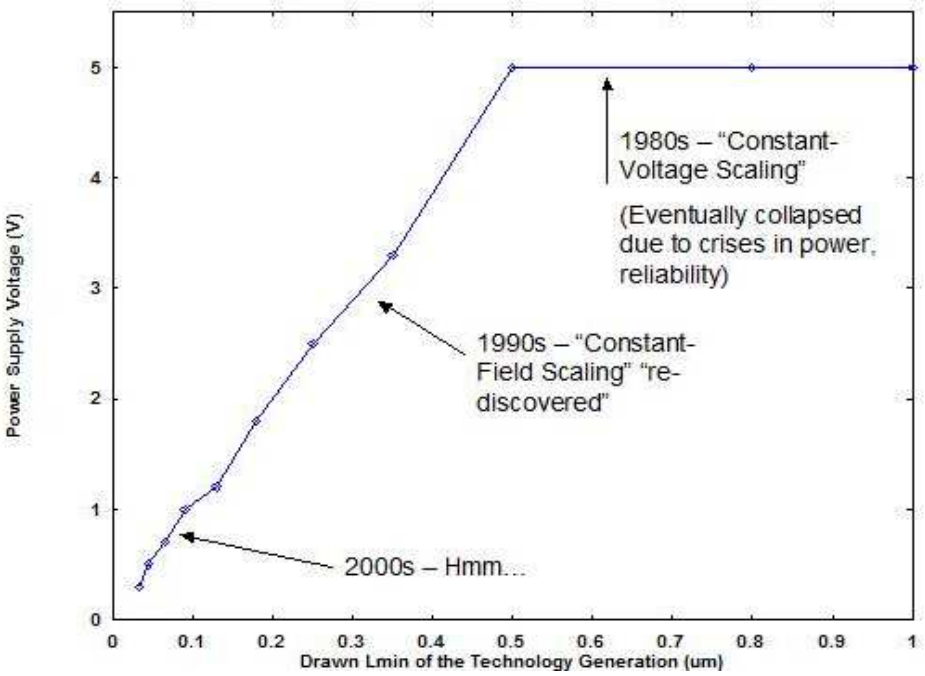


**Fig. 2.** The evolution of the power supply voltage over various technology generations, showing how constant-field scaling was "rediscovered" to replace the original constant-voltage scaling; the evolution to sub-100 nm dimensions, however, produced various difficulties.

However, there was another subtle crisis brewing – one which had been noted many years back [11], [12]; the MOS transistor threshold voltage is unable to scale in proportion to the scaled reduction of the power supply voltage. This is illustrated in Figure 3.

The original Mead concept of constant-field scaling requires the threshold voltage to scale in proportion to the power supply voltage; however, this cannot happen because the threshold voltage is limited by the thermal voltage $kT/q$.

In reality, the original Mead scaling idea was actually incomplete – and needs to be updated and amended as follows: *As the dimension goes to the limit of zero and the voltage goes to the limit of zero with it, the absolute temperature should* ***also*** *go to zero with both of them.*

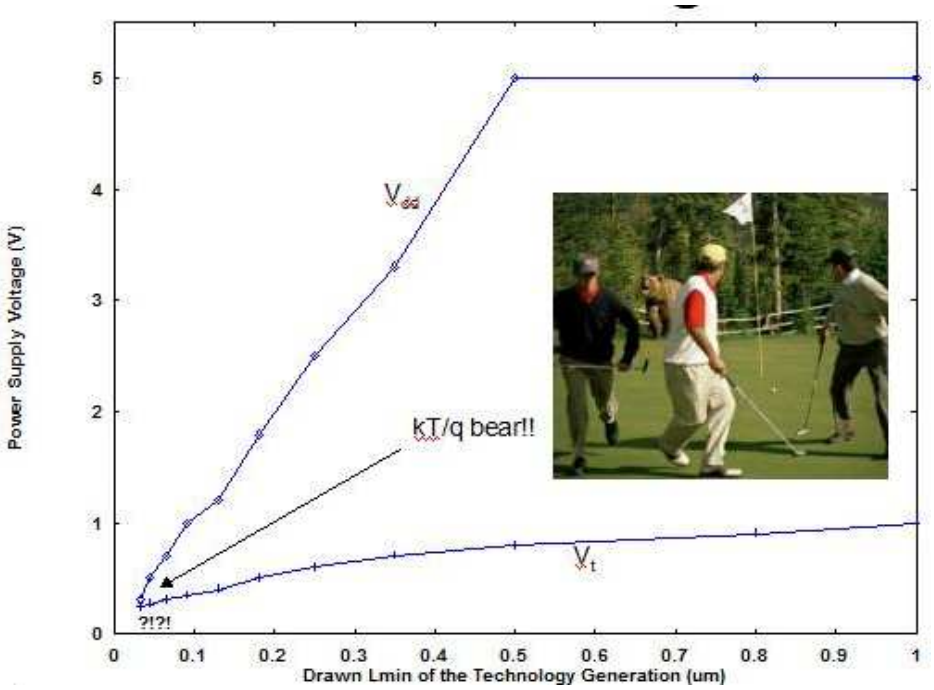


**Fig. 3.** The evolution of the power supply voltage along with the evolution of the threshold voltage; the non-scaling nature of the threshold voltage, due to the $kT/q$ granularity, is particularly frightening.

This isn't a statement of ideology or intent, but of the basic nature of semiconductor physics. However, the biggest challenge posed by this reality is partly tangible, but in fact more intellectual – *all of the extant long-used methods, approaches, and "understandings" are based on an assumption of the power supply voltage being considerably larger than the threshold voltage.* The problem is that reality is trying to tell us that those methods are obsolete and must be replaced. Thus, the conundrum that has emerged in recent years has been as follows. Do we discard constant-field scaling again? Or do we develop new methods to address the new situation of emerging granularities?

Unfortunately, the chosen engineering solution has been to try to avoid the granularities by discarding constant-field scaling again; this is depicted in Figure 4.

As technology dimensions dropped below 100 nm, the power supply voltage – which was originally projected to properly follow the dimensions and drop below 1 V – has instead been pegged at 1 V. From Figure 4, it is clear that this sub-100 nm situation is reprising what happened earlier when constant-voltage scaling pegged the supply voltage at 5 V – and the same dolorous consequences are appearing once again. Later, an interesting method of analysis will be used to show in frighteningly vivid detail why "traditional" design methods collapse completely for supply voltages less than 1 V.

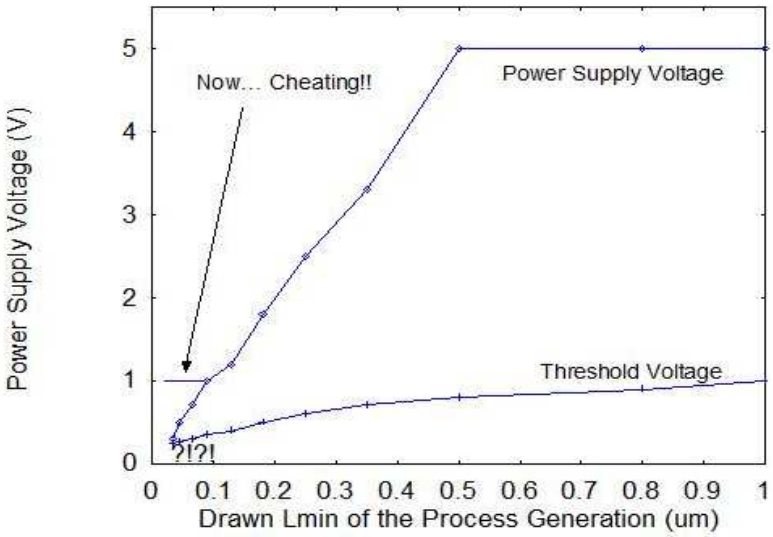


**Fig. 4.** An "updated" version of the evolution of the CMOS power supply voltage; due to the inability to scale the threshold voltage, below 100 nm the power supply has been fixed at 1 V – thus reintroducing constant voltage scaling.

However, we can distill the thrall that outdated thinking has had by examining the situation in analog/RF CMOS design. Digital design, being coarse and driven by its own particulars, could suitably rely on a notion of the MOSFET based entirely on its current and the loose idea of a "threshold voltage." Decades back, the basic methods of analog/RF CMOS design adopted the same lines of attack, even though they really aren't proper – in analog/RF applications, the MOS transistor is not being used as a switch; instead, it is being used as a charge-tuned signal-response device.

As dimensions decreased and power supply voltages came down, analog/RF faced a crisis of its own – a crisis due mostly to the now-outdated methods that rely upon the assumption that the power supply voltage was much larger than the threshold voltage. However, the more-contemporary reality is that this design "headroom" has just about disappeared; once again, this is a consequence of the inability to scale the threshold voltage due to the fixed-in-place nature of the thermal voltage $kT/q$.

A superior (and more modern) approach to describing the MOS transistor via its charge behavior *and of unifying that description directly with design applications* is

to describe the MOS transistor in terms of an inversion coefficient, which serves as a suitable doppelganger for the channel charge. This approach transcends outdated and coarse notions of a "threshold voltage," treats the MOS transistor (properly) as being of continuous operation over a large range of inversion conditions, and provides a unified description that is essentially independent of the channel length. This description is outlined in Figure 5.

In weak inversion, the transconductance efficiency ($g_m/I_d$) goes to the thermal voltage $kT/q$. The rest of the device behavior, independent of geometry, is described across a continuum of channel inversion charge densities – and it is the specific channel inversion charge density that provides the key guidepost for analog/RF design usage. In strong inversion, there is geometry dependence to the behavior, as the MOS transistor is bounded by two asymptotic limits. For long channel devices, the square-law asymptote exists – *but is (and* ***never can be****) reached*. To reach the square-law limit, the transistor would have to become "large" in three dimensions rather than two – that is, it would have to become "large" in the vertical dimension rather than in just the lateral (width and length) dimensions. In the short channel limit, the MOS transistor behavior goes to a linear limit. Thus, the description of Figure 5 provides a very simple yet very complete description of the modern MOS transistor – a description that is much more consonant with the notion of the MOS transistor as the first practical nanoscale electron device.

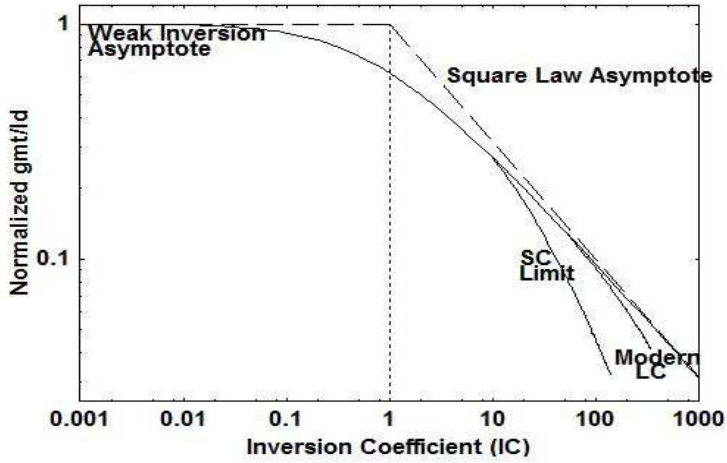


**Fig. 5.** A more modern method of interpreting the MOS transistor – one that transcends the "traditional" reliance on gate voltage headroom, and that is consonant with use in analog/RF design.

A measured example of this method is depicted in Figure 6. This is a measured set of data for various channel lengths from a 0.18 μm process. As described above, here we see a single, unified description across the entire spectrum of charge response. In addition, the characteristics separate in strong inversion, with the short (minimum channel length) device going to the linear limit, and the longest channel device not able to go to the square-law limit.

A set of measured data in the form of Figure 6 is presented in Figure 7. This data set is for several channel lengths, but from a 0.5 μm process. The form of the results shown in Figure 7 is the same as that shown in Figure 6 – *even though the two data sets are from completely different processes, completely different process generations, and completely different manufacturers.* Thus, we can see that this description is universal across processes and process generations. Further, we can note that this description – by being universal – is clearly useful for "beyond-MOS" nanoscale electron device families.

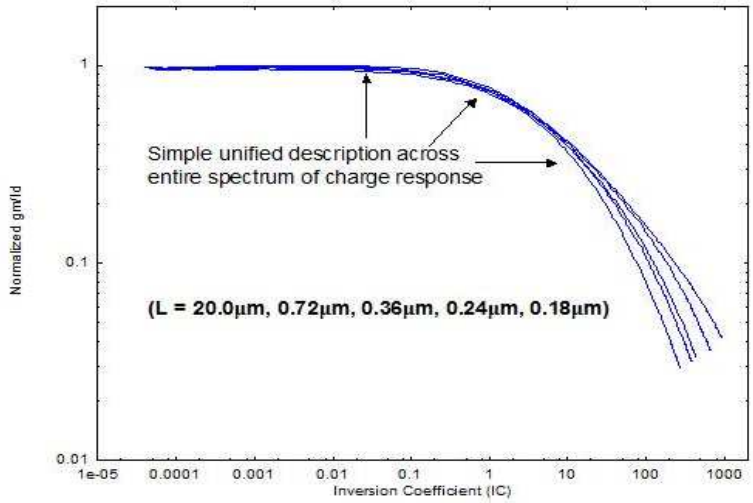


**Fig. 6.** Measured data for several different channel lengths in a 0.18 μm process, demonstrating experimentally the theoretical contentions shown earlier.

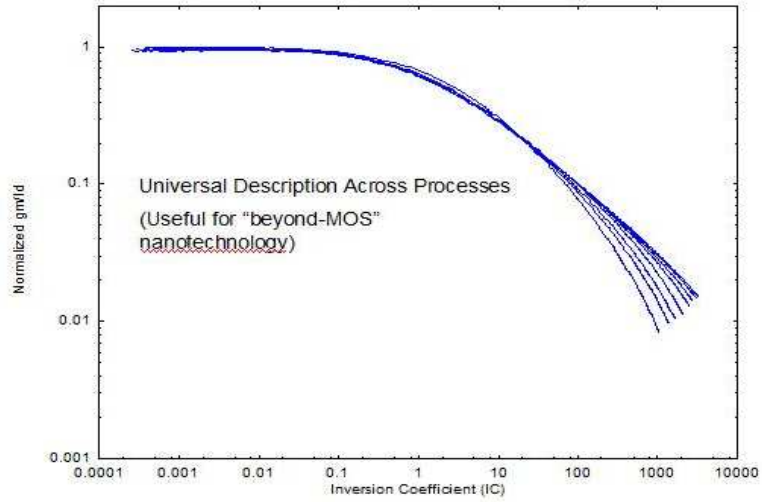


**Fig. 7.** Measured data for several different channel lengths in a 0.5 μm process, demonstrating the universal nature of this method.

We can now use the method and metrics described above to more carefully examine what happens in the MOS transistor when the inability to scale the threshold voltage induces the severe loss of headroom. This is done by examining the characteristics for the shortest channel device over several process generations, as depicted in Figure 8. This chart is actually rather horrifying, as it tells us that under the present regime of MOS scaling, *we are losing our ability to strongly invert the MOS transistor.* In the analog/RF domain in particular, the capacity for bandwidth is very sensitive to how strongly one is able to drive the MOS transistor into strong inversion – and as Figure 8 clearly shows, *the reduction of geometries is actually not providing more bandwidth.* This indeed does appear to be the case, as for all practical purposes RF-CMOS has been stalled at a working limit of about 5 GHz.

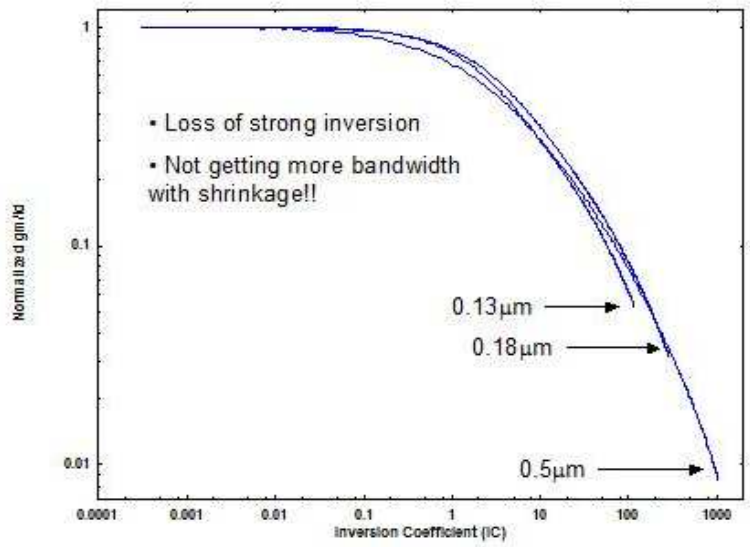


**Fig. 8.** Measured data for the minimum channel length device in several different process generations, showing how the loss of headroom is causing a severe (and bandwidth-crushing) loss in the ability to strongly invert the MOS transistor in more-scaled process generations.

The results of Figure 8 can be converted into an examination of how the maximal ability to drive the minimum-geometry transistor behaves over process generations; this is shown in Figure 9.

"Traditional" analog/RF CMOS design methods face a *de facto* self-imposed minimum inversion coefficient limit of about 25 – a lower limit largely self-imposed by outdated thinking and outdated methods. At the same time, the maximum usable inversion coefficient has been dropping very sharply – causing the obvious squeeze depicted in Figure 9. Thus, Figure 9 turns out to have made a surprisingly-accurate technology prediction – that "traditional" analog/RF CMOS design would collapse completely below 100 nm (requiring its replacement with something more appropriate), or the power supply voltage would no longer be scaled so as to retain the small

remaining amount of "headroom" shown in Figure 9. Not surprisingly, the inertial choice prevailed – and as Figure 4 shows, a conscious decision has been made to return to constant-voltage scaling (this time at 1 V) for sub-100 nm process technologies.

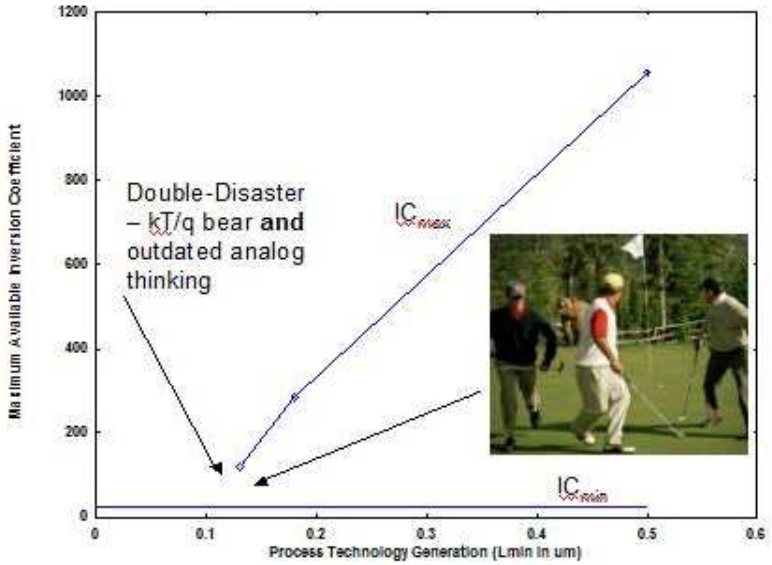


**Fig. 9.** The maximum level of ability to invert the MOS transistor as process generations scale; clearly, in the absence of a change, "traditional" methods of analog/RF design collapse completely below 100 nm.

As a brief concluding comment for this section, it should also be noted that the "traditional" design methods, being based on inappropriate foundations, do not provide a clean method of closing up analog/RF designs. Instead, the little-discussed engineering practice is to use extensive iterations – in both design and (more frighteningly) in silicon. This embarrassing shortcoming is little-discussed in the engineering literature, but can be discerned by careful analysis of required filings with financial industry regulators.

## 6. Diversion – Side Implications

Earlier it was noted that the continued down-scaling of MOS technology has collided with the granularity of the thermal voltage $kT/q$. It was also noted that this is due to the incompleteness of the original Mead scaling insight – that in reality, as the dimension and the voltage go to the limit of zero, the absolute temperature should go to zero with them.

The slightly strange implication from this reality is that *the only way to maintain "traditional" approaches to MOS technology is to scale the temperature.* That is, *true* scaling requires a concomitant proportional reduction in the absolute operating temperature.

Scaling the temperature is the only proper way to scale the threshold voltage, as shown in Figure 10. Reducing the absolute temperature to a quarter of its room temperature value allows for a similar reduction in the threshold voltage.

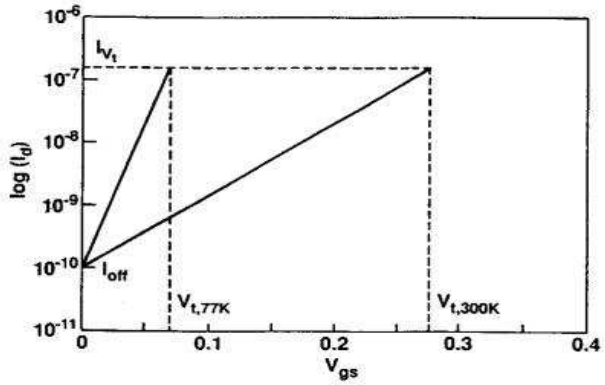


**Fig. 10.** The design of the threshold voltage, based on an off-current goal and the $kT/q$-determined subthreshold slope; only by reducing the temperature can the threshold voltage truly be scaled.

In a similar fashion, the power supply voltage should also be reduced by the same temperature-induced proportion. For example, for a 0.18 μm process, the standard supply voltage of 1.8 V should scale (under this scheme) to a quarter of that value – or 0.45 V. Thus, the turn-on characteristics for 0.18 μm devices – designed for 300 K and 77 K – are depicted in Figure 11.

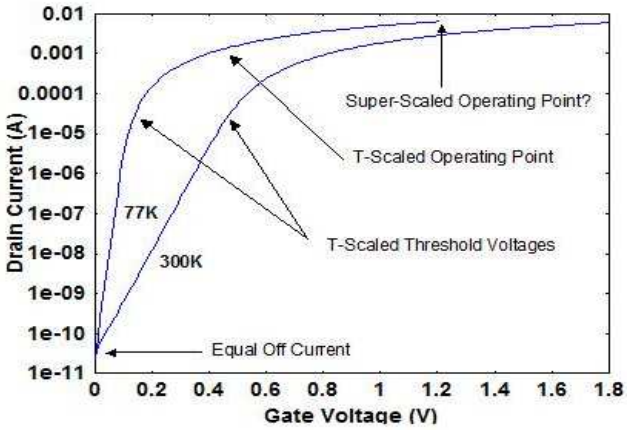


**Fig. 11.** The turn-on characteristics for 0.18 μm devices, where the temperature-based scaling of the threshold voltage has been properly exploited; the reliability constraints imply that the low temperature power supply voltage can be made higher than "pure" scaling would deem.

The operating point of 0.45 V is reached by the pure application of scaling theory. However, in reality, the upper limit on the power supply voltage is actually set by hot-

carrier reliability – and in this case, the power supply voltage can in fact be higher, at 1.2 V; this is also shown in Figure 11. Note how this super-scaled operating point provides a great deal more proportionate "headroom" than is possible for the room-temperature operating point. This situation is clarified in Figure 12.

The super-scaled operating point has indeed transcended the squeeze that has been placed on the maximum inversion coefficient – easily reaching inversion values similar to those of the "good old days" in analog/RF MOS design. Careful examination of the 77K curve even indicates that the linear limit associated with the minimum channel length has *not* in fact been reached – indicating that the further channel length reductions are possible.

The key point of this discussion has *not* been specific technology advocacy – as refrigeration is obviously not a desirable requirement to impose. The purpose was illustrative – to simply look at what the pure physics tells us we should be doing. *One must either agree to follow the physics – or find some way to do things in a completely different manner.* The inversion-coefficient-based approach to understanding and interpreting the MOS transistor – and then using that quantified approach in analog/RF CMOS design – presents a unique way of transcending the traditional methods and limits.

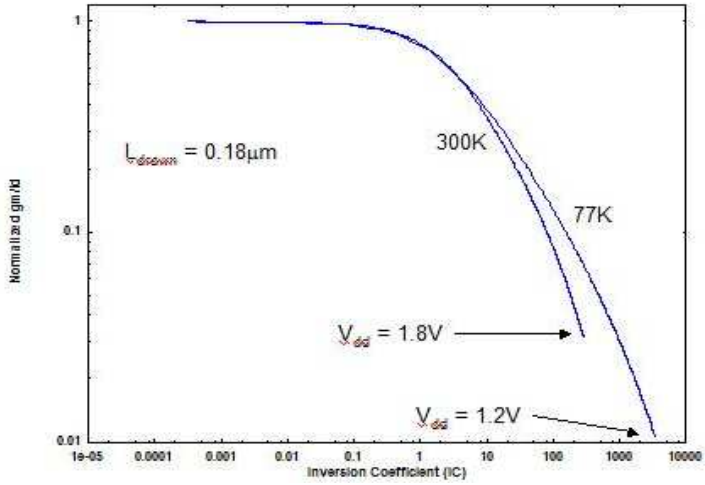


**Fig. 12.** A comparison of the ability to invert 0.18 µm devices at the two operating temperatures; the use of low temperature to scale the threshold voltage allows the low temperature device to be inverted to much larger levels – levels lost due to the $kT/q$ squeeze.

This discussion also tells us something particularly important about putative nanoscale electron devices. *As the "physics" collides with various "smallnesses" and granularities, new methods and new thinking will absolutely be required for the new situation.* It is a bonus that this need for new thinking also works backwards into sub-100-nanometer CMOS technology – particularly into issues of the analog/RF variety.

## 7. Nanoscale Electron Devices and RF Applications

We can now look beyond CMOS for further insights into nanoscale electron devices – particularly in the more demanding sets of applications. A good template for this analysis is the application of electron devices to high-end RF applications – particularly at the edge of the frequency frontier.

Unfortunately, a great deal of "ideology" has developed around the notion of using CMOS for RF applications – and remarkably this "ideology" has persisted against outcomes for more than a decade. In terms of electron devices, bipolar devices provide a competitive and viable option – indicating that the bipolar transistor will be a major player in the game of nanoscale electron devices.

There are three critical issues regarding RF-CMOS that must be understood.

First, CMOS is *not* the low-power option in RF applications. The low-power nature of CMOS is based on the capacitive coupling between the input signal and the output action; when CMOS is operated in a quasi-static (digital) fashion, the signals on the MOS gate are effectively DC. Of course, at DC the capacitor is an open circuit – and thus the power consumption is low. However, in the RF domain, the signals applied to the MOS gate are AC; in the AC realm, the MOS gate is not an open but becomes a *short* – and becomes a more perfect short at higher frequencies. Thus, the MOS transistor provides an AC path to ground – a depicted simply in Figure 13.

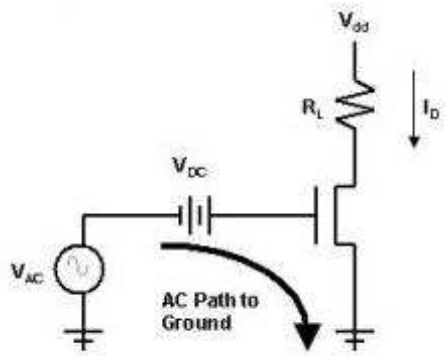


**Fig. 13.** A simply MOS circuit, showing how there is an AC path to ground.

Thus, there is substantial power consumption – something that can actually be verified in extant system products.

Second, CMOS is *not* the lowest-cost choice. Comparisons of cost cannot be done at "equal lithography" – but must be done at "equal performance." Several extant silicon-germanium bipolar/BiCMOS processes actually provide much lower cost for the same performance level – and (as noted earlier) for any application above about 5GHz, CMOS is simply (in practical terms) incapable of delivering.

Third, complete integration is *not* a desirable RF design choice. Complete integration is a handy concept in digital design, when all the parts are similar in content

and signal types – and "play nice" together. In contrast, digital switching noise is very detrimental to the behavior of nearby analog and RF circuitry – keeping the circuits sufficiently separate is a better option. Yield is the product of independent systems – and so one independent failure in a fully-integrated system takes the whole things down with it. In addition, full integration places intolerable constraints (both technical *and* financial) among the incompatible (RF, analog, digital) pieces.

The demands being placed on RF technology are excruciating – and that situation is becoming even more difficult. There is a need for higher data rates – but with very low power consumption and good transmission ranges. This situation can be graphically summarized via the "Iron Triangle" of wireless data communications, shown in Figure 14.

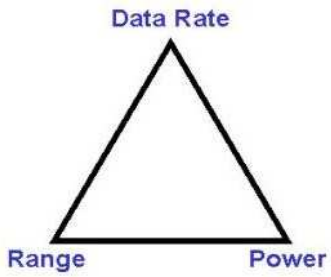


**Fig. 14.** The "Iron Triangle" of wireless data communications, indicating that the three key factors that naturally trade off against each other; in the absence of some way of increasing the trade-off space, one of these factors can be improved only by harming one (or both) of the other factors.

This description is actually consonant with basic information theory; data rate, range, and power consumption are engineering factors *that trade off directly against each other*. In other words, for a given situation, one of these factors can be improved *only by paying a price in one (or both) of the other factors*. The only way out of this restriction is to find some exogenous way to increase the size of the triangle – and thus improve the overall trade-off situation.

Given its subpar power-efficiency, CMOS is actually a poor choice in this situation – and again, this is demonstrated by comparative analysis of available products. Because of this, compound-semiconductor-based power amplifiers continue to flourish – despite their high cost and low integration, they simply provide very good power efficiencies. Most mobile telephone power amplifiers have used gallium arsenide – and this situation holds today. More recently, the extremely good power efficiency of gallium nitride power amplifiers has induced their deployment in low-volume applications, such as base stations – where the expense is worth the efficiency. Compound semiconductors will remain important in RF microelectronics; only silicon-germanium bipolar technology offers competitive abilities in power efficiency, while CMOS does not.

## 8. From Electron Devices to Systems – Not a Choice, But a Necessity

We now come to a final issue that represents a sea change in our thinking about electronic systems – one that has gone largely unrecognized. Much of the success of semiconductor electronics – and systems built on semiconductor electronics – has been based on the ability to separate the tasks into very-forgiving sub-sections, work independently on those sub-sections, and then integrate things together with little worry.

That era, however, is coming to an end. The core problem is that the sub-sections can no longer be regarded as "linearly independent." In other words, *everything interacts so strongly with everything else that those interactions are fundamental to the design and development of everything in the complete system.* This indeed means that everything is now in fact system design.

A good example of this change is again found by examining the situation in RF microelectronics. In general, integrated circuit design has been regarded as being "silicon-based" – if a "chip" worked as it was on the wafer, then success had definitely been achieved, with packaging as merely an afterthought. However, in RF microelectronics, packaging introduces a major perturbation into the system – working integrated circuits too often cease to work properly when they are packaged. The implications are obvious, but disturbing to entrenched thinking – *packaging is no longer an afterthought, but is an integral part of the design.* The same notion applies to the further perturbations that are introduced by the deployment into the final system.

Again, *all* the pieces of the puzzle must be co-designed together to deliver proper system behavior. *Of particular note, overall engineering trade-offs* ***cannot*** *be hermetically isolated in the individual tasks; instead, they must be* ***globally amortized*** *over the entire system.* In contrast, efforts to declare "final" results at small-scale levels are of increasingly questionable value.

In a larger sense, this observation points us to another emerging factor in nanoscale electron devices and nanotechnology-based systems. In the past, most of the focus has been on the *components themselves* – rather than on *how those components all go together.*

Much of this paper has dealt with various aspects of high-frequency integrated-circuit design, and with good reason – as noted intermittently, a proper respect for this challenge bridges the gap between mere components and the larger picture of systems.

A key neglected (and missing) factor in RF microelectronics has been an appreciation – that how the pieces interact is more important than the pieces themselves. This indicates that nanoscale electron devices and nanotechnology-based systems will represent a fundamental intellectual break with the past – in that they will require that network concepts and system constraints be included at very fundamental levels.

Ultimately, by building in these network principles, "emergent behavior" will occur – which is a fancy way of saying that complex network behavior, such as that of neural networks, will appear on its own.

At the higher levels, new network topologies are desperately needed; nanotechnology devices will (and must) incorporate network ideas at the most fundamental levels. These network ideas should actually fit naturally into nanotechnology as it fully begins to emerge.

## 9. Conclusions

This paper has considered some of the large-scale issues involved in the development of nanoscale electron devices. It was noted that most of the challenges are intellectual rather than material.

It was noted that the prime challenge for nanotechnology is the identification of a particular abundance/scarcity intersection where it can be decisive. It was also noted that rather than just bulldozing ahead with a notion of "smaller and smaller" – that is, the sequential and (low-thought-content) continual application of "scale" – the focus must be on new thinking, new knowledge, and the new methods that flow from them. Consideration was given to the various granularities that are emerging – granularities that are material, as well as those that involve fundamental physics, such as the critical importance of the thermal voltage $kT/q$.

It was suggested that the major emerging need is demand for much more bandwidth. Present methods attempt to generate bandwidth in the old-fashioned "digital" way, by throwing power and transistors at the problem – and these methods are simply not delivering. Of particular note is a nanotechnology success that is driving this situation even further into need – the incredible growth in storage, and the ability to distribute that storage among a galaxy of portable devices; as storage rises rapidly as a fundamental abundance, the situation demands much more wireless bandwidth, and at very low power consumption – something which the old-fashioned methods are grossly unable to address.

Finally, it was noted that nanotechnology needs to find a cost breakthrough of its very own – one that will make nanotechnology truly indispensable in the manner that CMOS is presently indispensable. In general, new network topologies are needed – and these can be (and must be) folded naturally into nanotechnology.

As a final flourish, it should be noted that to make "small" work, we must **think big**.

# References

[1] See, for example, *Nanotech Buzz* (`http://www.nanotechbuzz.com`), 26 January 2006.

[2] IEEE EDS/SSCS Mini-Colloquium, *Nanoelectronic Devices – Present and Perspective*, Sinaia, Romania, 14 October 2007.

[3] FOTY D., *Electron Devices in Nanoelectronic Circuits and Systems: Problems and Prospects*, IEEE EDS/SSCS Mini-Colloquium, *Nanoelectronic Devices – Present and Perspective*, Sinaia, Romania, 14 October 2007.

[4] SINGH R., *Challenges and Opportunities in the Fields of Information Processing and Energy Conversion in the Nano World of the* $21^{st}$ *Century*, IEEE EDS/SSCS Mini-

Colloquium, *Nanoelectronic Devices – Present and Perspective*, Sinaia, Romania, 14 October 2007.

[5] GILDER G., *Wired*(magazine), December 1996.

[6] BECKMANN P., *A History of* $\pi$, New York: St. Martin's Press, 1971.

[7] LINCOLN A., Annual Message to Congress, 1 December 1862.

[8] MEYYAPPAN M., *The Role of Nanotechnology in Shaping Nanoelectronics: An Overview*, IEEE EDS/SSCS Mini-Colloquium, *Nanoelectronic Devices – Present and Perspective*, Sinaia, Romania, 14 October 2007.

[9] WILD A., *'More Moore' Versus 'More than Moore': Status and Perspective*, IEEE EDS/SSCS Mini-Colloquium, *Nanoelectronic Devices – Present and Perspective*, Sinaia, Romania, 14 October 2007.

[10] HOENEISEN B., MEAD C., *Fundamental Limitations in Microelectronics I: MOS Technology*, Solid-State Electronics, vol. **15**, pp. 819–829, 1972.

[11] NOWAK E., *Ultimate CMOS ULSI Performance*, Digest of Technical Papers, 1993 International Electron Devices Meeting (IEDM), pp. 115–118.

[12] FOTY D., NOWAK E., P*erformance, Reliability, and Supply Voltage Reduction, with the Addition of Temperature as a Design Variable*, Proceedings of the 1993 European Solid State Device Research Conference (ESSDERC), pp. 943–948.